Выполнил студент БВТ2002 Митрофанов Алексей

Открытый курс по машинному обучению.

Тема 1. Первичный анализ данных с Pandas

Практическое задание. Анализ данных пассажиров "Титаника"

**Заполните код в клетках (где написано "Ваш код здесь")

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import math as m
%matplotlib inline
```

Считаем данные из файла в память в виде объекта Pandas.DataFrame

In [ ]:
```
data = pd.read_csv('titanic_train.csv',
                   index_col='PassengerId')
```

Данные представлены в виде таблицы. Посмотрим на первые 5 строк:

In [ ]:
```
data.head(5)
```

Out[ ]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	

In [ ]:
```
data.describe()
```

Out[ ]:

	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Для примера отберем пассажиров, которые сели в Cherbourg (Embarked=C) и заплатили более 200 у.е. за билет (fare > 200).

Убедитесь, что Вы понимаете, как эта конструкция работает.
Если нет – посмотрите, как вычисляется выражение в квадратных в скобках.

In [ ]:
```
data[(data['Embarked'] == 'C') & (data.Fare > 200)].head()
```

Out[ ]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
119	0	1	Baxter, Mr. Quigg Edmond	male	24.0	0	1	PC 17558	247.5208
259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3292
300	1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50.0	0	1	PC 17558	247.5208
312	1	1	Ryerson, Miss. Emily Borie	female	18.0	2	2	PC 17608	262.3750
378	0	1	Widener, Mr. Harry Elkins	male	27.0	0	2	113503	211.5000

Можно отсортировать этих людей по убыванию платы за билет.

In [ ]:
```
data[(data['Embarked'] == 'C') &
     (data['Fare'] > 200)].sort_values(by='Fare',
                                       ascending=False).head()
```

Out[ ]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3292	NaN
680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	0	1	PC 17755	512.3292	B51 B53 B55
738	1	1	Lesurer, Mr. Gustave J	male	35.0	0	0	PC 17755	512.3292	B101
312	1	1	Ryerson, Miss. Emily Borie	female	18.0	2	2	PC 17608	262.3750	B57 B59 B63 B66
743	1	1	Ryerson, Miss. Susan Parker "Suzette"	female	21.0	2	2	PC 17608	262.3750	B57 B59 B63 B66

Пример создания признака.

In [ ]:
```
def age_category(age):
    '''
    < 30 -> 1
    >= 30, <55 -> 2
    >= 55 -> 3
    '''
    if age < 30:
        return 1
    elif age < 55:
        return 2
    else:
        return 3
```

In [ ]:
```
age_categories = [age_category(age) for age in data.Age]
```

In [ ]:
```
data['Age_category'] = age_categories
```

Другой способ – через `apply`.

In [ ]:
```
data['Age_category'] = data['Age'].apply(age_category)
```

In [ ]:
```
data['Age_category'].value_counts()
```

Out[ ]:
```
1    384
2    288
3    219
Name: Age_category, dtype: int64
```

1. Сколько мужчин / женщин находилось на борту?

- 412 мужчин и 479 женщин
- 314 мужчин и 577 женщин
- 479 мужчин и 412 женщин
- * **`577 мужчин и 314 женщин`** * ✅

In [ ]:
```
data['Sex'].value_counts()
```

Out[ ]:
```
male      577
female    314
Name: Sex, dtype: int64
```

2. Выведите распределение переменной `Pclass` (социально-экономический статус) и это же распределение, только для мужчин / женщин по отдельности. Сколько было мужчин 2-го класса?

- 104
- * **`108`** * ✅
- 112
- 125

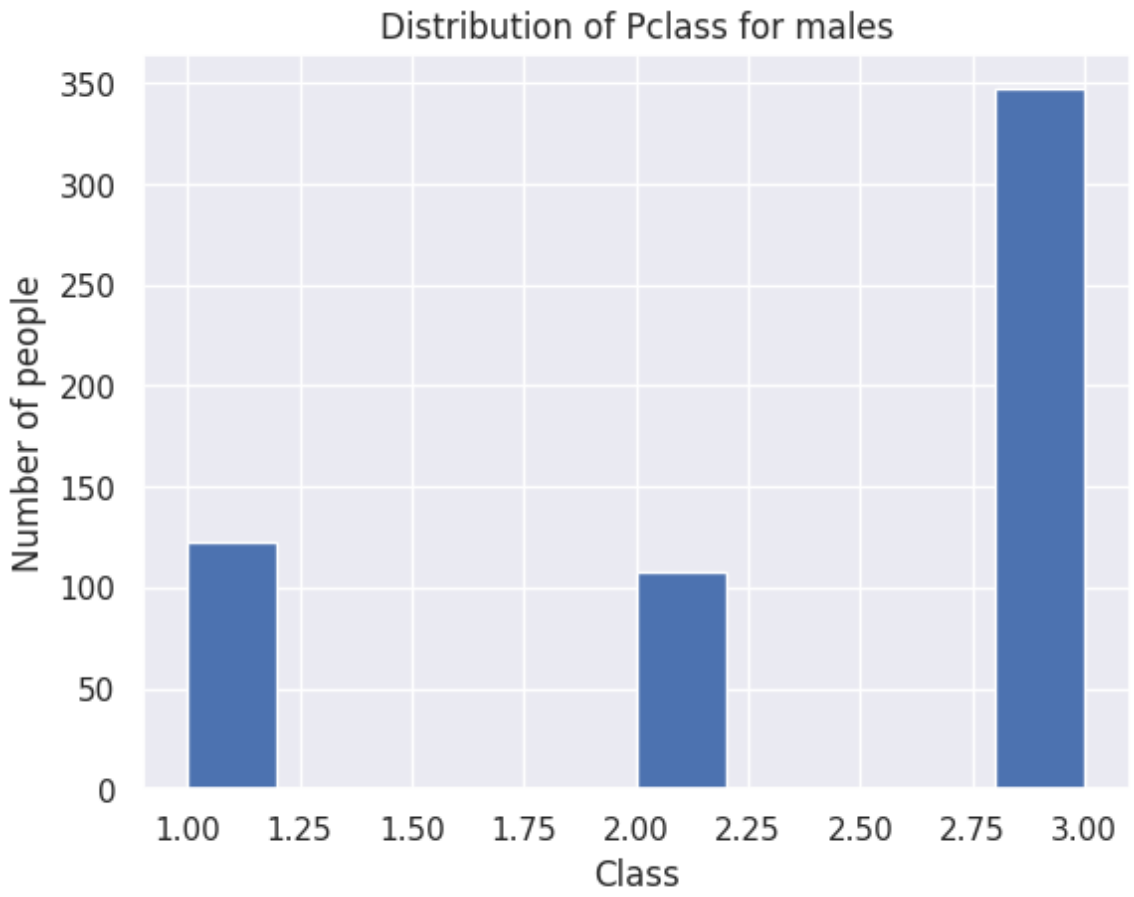
In [ ]:
```
import seaborn as sns
sns.set()
```

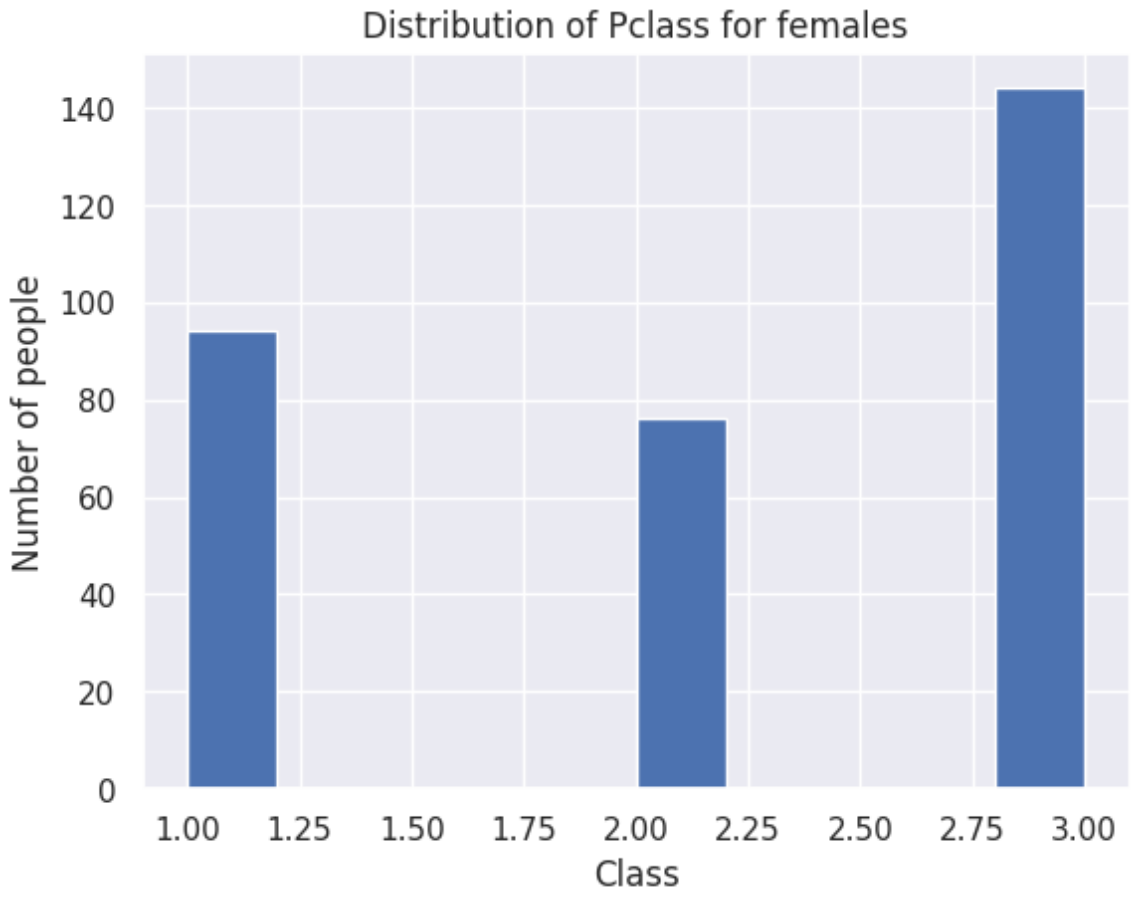
In [ ]:

```
# Ваш код здесь
data.hist(column='Pclass')
ax = plt.gcf().gca()
ax.set_title('General distribution of Pclass')
ax.set_xlabel('Class')
ax.set_ylabel('Number of people')

data[data['Sex'] == 'male'].hist(column='Pclass')
ax = plt.gcf().gca()
ax.set_title('Distribution of Pclass for males')
ax.set_xlabel('Class')
ax.set_ylabel('Number of people')

data[data['Sex'] == 'female'].hist(column='Pclass')
ax = plt.gcf().gca()
ax.set_title('Distribution of Pclass for females')
ax.set_xlabel('Class')
ax.set_ylabel('Number of people')
plt.show()
```





In [ ]:
```
print( "Males of Pclass 2:", data[(data['Sex'] == 'male') & (data['Pclass
```

```
Males of Pclass 2: 108
```

3. Каковы медиана и стандартное отклонение платежей (`Fare`)? Округлите до 2 десятичных знаков.

- * **`Медиана - 14.45, стандартное отклонение - 49.69`** * ✅
- Медиана – 15.1, стандартное отклонение – 12.15
- Медиана – 13.15, стандартное отклонение – 35.3
- Медиана – 17.43, стандартное отклонение – 39.1

In [ ]:
```
print("Median: ", round(data['Fare'].median(), 2))
print("Std. deviation: ", round(data['Fare'].std(),2))
```

```
Median:  14.45
Std. deviation:  49.69
```

4. Правда ли, что люди моложе 30 лет выживали чаще, чем люди старше 60 лет? Каковы доли выживших в обеих группах?

- 22.7% среди молодых и 40.6% среди старых
- * **`40.6% среди молодых и 22.7% среди старых`** * ✅
- 35.3% среди молодых и 27.4% среди старых
- 27.4% среди молодых и 35.3% среди старых

In [ ]:
```
# Ваш код здесь

under_30_survived = data[(data['Age'] < 30) & (data['Survived'] == 1)].sha
over_60_survived = data[(data['Age'] > 60) & (data['Survived'] == 1)].shap

under_30_total = data[data['Age'] < 30].shape[0]
over_60_total = data[data['Age'] > 60].shape[0]

print("Percentage of survivors under 30: ", round(under_30_survived / (und
print("Percentage of survivors over 60: ", round(over_60_survived / (over_
```

```
Percentage of survivors under 30:  40.6 %
Percentage of survivors over 60:  22.7 %
```

5. Правда ли, что женщины выживали чаще мужчин? Каковы доли выживших в обеих группах?

- 30.2% среди мужчин и 46.2% среди женщин
- 35.7% среди мужчин и 74.2% среди женщин
- 21.1% среди мужчин и 46.2% среди женщин
- * **`18.9% среди мужчин и 74.2% среди женщин`** *

In [ ]:
```
# Ваш код здесь

males_survived = data[(data['Sex'] == 'male') & (data['Survived'] == 1)].s
females_survived = data[(data['Sex'] == 'female') & (data['Survived'] == 1

males_total = data[(data['Sex'] == 'male')].shape[0]
females_total = data[(data['Sex'] == 'female')].shape[0]

print('Males survived: ', round(males_survived / (males_total / 100), 1),
print('Females survived:', round(females_survived / (females_total / 100)
```

```
Males survived:  18.9 %
Females survived: 74.2 %
```

6. Найдите самое популярное имя среди пассажиров Титаника мужского пола?

- Charles
- Thomas
- * **William** *
- John

In [ ]:
```
# Ваш код здесь

def first_name(name):
    import re
    # регулярка, ищущая первое слово, идущее сразу после 'Mr.' или 'Mrs.'
    return re.search(r'(?<=\b[Mr\.|Mrs\.]\s)(\w+)|$', name).group()


data['First_name'] = data['Name'].apply(first_name)

data[data['Sex'] == 'male']['First_name'].value_counts()
```

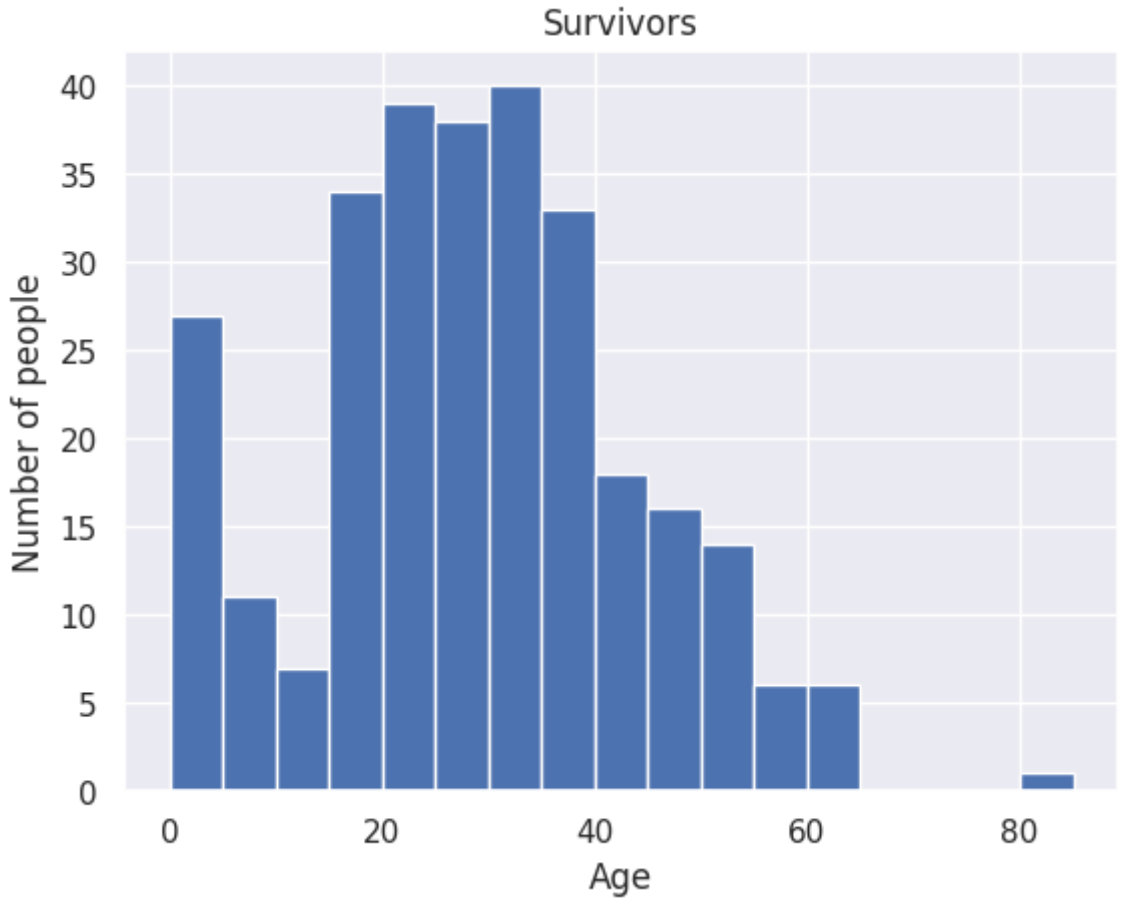
Out[ ]:
```
William    35
John       25
George     14
Charles    13
Thomas     13
           ..
Ignjac      1
Yoto        1
Austen      1
Mitto       1
Juozas      1
Name: First_name, Length: 286, dtype: int64
```

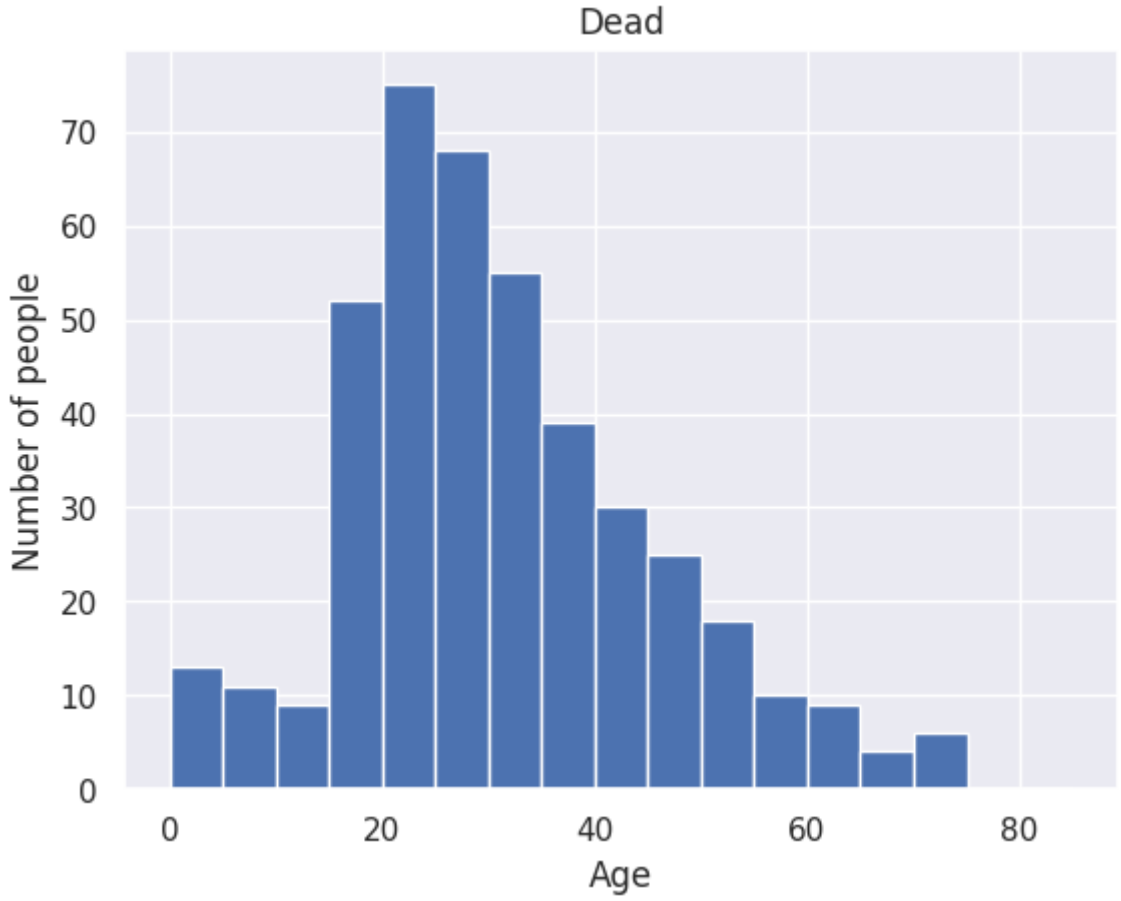
7. Сравните графически распределение стоимости билетов и возраста у спасенных и у погибших. Средний возраст погибших выше, верно?

- * **Да** * ✅
- Нет

In [ ]:
```
data[data['Survived'] == 1].hist(column='Age', bins=range(0, 90, 5))
ax = plt.gcf().gca()
ax.set_title('Survivors')
ax.set_xlabel('Age')
ax.set_ylabel('Number of people')

data[data['Survived'] == 0].hist(column='Age', bins=range(0, 90, 5))
ax = plt.gcf().gca()
ax.set_title('Dead')
ax.set_xlabel('Age')
ax.set_ylabel('Number of people')
plt.show()
```





In [ ]:
```
# проверка среднего возраста
data[data['Survived'] == 1]['Age'].mean()
```

Out[ ]:
```
28.343689655172415
```

In [ ]:
```
data[data['Survived'] == 0]['Age'].mean()
```

Out[ ]:
```
30.62617924528302
```

- **Средний возраст погибших выше** из-за большого числа выживших детей

8. Как отличается средний возраст мужчин / женщин в зависимости от класса обслуживания? Выберите верные утверждения:

- В среднем мужчины 1-го класса старше 40 лет ✅
- В среднем женщины 1-го класса старше 40 лет ❌
- Мужчины всех классов в среднем старше женщин того же класса ✅
- В среднем люди в 1 классе старше, чем во 2-ом, а те старше представителей 3-го класса ✅

In [ ]:
```
# Ваш код здесь

for pclass in range(1, 4):
    male_avg_age = data[(data['Sex'] == 'male') & (data['Pclass'] == pclas
    female_avg_age = data[(data['Sex'] == 'female') & (data['Pclass'] == p
    print(f'Male avg. age in Pclass {pclass}: \t{male_avg_age}')
    print(f'Female avg. age in Pclass {pclass}: \t{female_avg_age}')
```

```
Male avg. age in Pclass 1:      41.28138613861386
Female avg. age in Pclass 1:    34.61176470588235
Male avg. age in Pclass 2:      30.74070707070707
Female avg. age in Pclass 2:    28.722972972972972
Male avg. age in Pclass 3:      26.507588932806325
Female avg. age in Pclass 3:    21.75
```