НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Дисциплина: «Анализ данных»

Домашнее задание на тему:

«Лабораторная работа №1»

Выполнил: Осипов Лев,

студент группы 301ПИ (1).

Москва, 2015 г.

**СОДЕРЖАНИЕ**

**Задание 13**

**Задание 23**

**Задание 33**

**Задание 44**

**Задание 54**

**Задание 64**

**Задание 75**

**Задание 85**

**Задание 95**

**Задание 106**

**Список литературы7**

**Текст программы8**

**ЗАДАНИЕ 1**

Для решения задания были посчитаны отношения количества представителей каждого пола к общему количеству пассажиров.

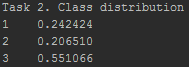


*Рис. 1*

По результатам программы (Рис. 1) видно, что мужчины составляли примерно 64,8% от общего числа пассажиров, тогда как женщины – примерно 35,2%.

**ЗАДАНИЕ 2**

Для решения задания были посчитаны отношения количества представителей каждого класса к общему количеству пассажиров.

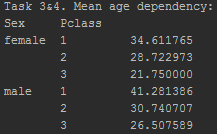
****

*Рис. 2*

По результатам программы (Рис. 2) видно, что в первом классе находилось примерно 24,2% пассажиров, во втором – 20,7%, а в третьем – 55,1%.

**ЗАДАНИЕ 3**

Для решения задания 3 и 4 были высчитаны количественные значения принадлежности каждого пола к каждому классу и взяты средние.

****

*Рис. 3*

По результатам программы (Рис. 3) видно, что в первом классе средний возраст у мужчин – 41,3 лет, во втором классе – 30,7 лет, а в третьем – 26,5 лет. Заметно, что с ростом престижности класса средний возраст также растет.

**ЗАДАНИЕ 4**

По результатам программы (Рис. 3) видно, что в первом классе средний возраст у женщин – 34,6 лет, во втором классе – 28,7 лет, а в третьем – 21,8 лет. Можно видеть аналогичную тенденцию (рост престижа – рост возраста).

**ЗАДАНИЕ 5**

Для решения задания были посчитано отношение выживших людей к общему количеству пассажиров.

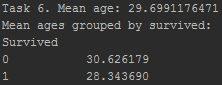
****

*Рис. 4*

По результатам программы (Рис. 4) видно, что выжили примерно 38,4% пассажиров.

**ЗАДАНИЕ 6**

Для решения задания был посчитан средний возраст всех пассажиров, а затем средние значение возрастов выживших и погибших пассажиров.

****

*Рис. 5*

По результатам программы (Рис. 5) видно, что средний возраст составлял примерно 29,7 лет. Средних возраст выживших – 28,3 лет, а погибших – 30,6 лет. Можно сделать вывод, что выживали действительно более молодые пассажиры.

**ЗАДАНИЕ 7**

Для решения задания были посчитаны отношения количества выживших представителей каждого пола к общему количеству пассажиров.

****

*Рис. 6*

По результатам программы (Рис. 6) видно, что выжило примерно 74,2% мужчин и 18,9% женщин. Можно сделать вывод, что чаще выживали мужчины.

**ЗАДАНИЕ 8**

Для решения задания было подсчитано среднее значение стоимости билетов, а для определения степени варьирования было подсчитано среднее абсолютное отклонение.

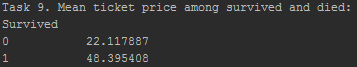
****

*Рис. 7*

По результатам программы (Рис. 7) видно, что средняя цена на билеты была примерно 32,2, а среднее абсолютное отклонение – примерно 28, 2.

**ЗАДАНИЕ 9**

Для решения задания были посчитано среднее значение стоимости билета для выживших и погибших.

****

*Рис. 8*

По результатам программы (Рис. 8) видно, что средняя цена билетов у выживших была примерно 48,4, а у погибших – 22,1. Можно сделать вывод, что действительно выживали пассажиры с более дорогими билетами. Это логично, потому что в приоритете спасения всегда элита и персонал ориентировался прежде всего на нее.

**ЗАДАНИЕ 10**

Для решения задания среди мужчин было взято самое встречающееся значения имени. Для фильтра префиксов было использовано регулярное выражение.

****

*Рис. 9*

По результатам программы (Рис. 9) видно, что самое популярное мужское имя на корабле – William.

**СПИСОК ЛИТЕРАТУРЫ**

1. **Анализ данных (Программная инженерия)** – http://wiki.cs.hse.ru/%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7\_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85\_%28%D0%9F%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%BD%D0%B0%D1%8F\_%D0%B8%D0%BD%D0%B6%D0%B5%D0%BD%D0%B5%D1%80%D0%B8%D1%8F%29#.D0.9E.D1.84.D0.BE.D1.80.D0.BC.D0.BB.D0.B5.D0.BD.D0.B8.D0.B5\_.D0.BF.D0.B8.D1.81.D0.B5.D0.BC

**ТЕКСТ ПРОГРАММЫ**

\_\_author\_\_ = 'Lev Osipov'  
  
**import** pandas **as** pd  
  
# Reading data file  
df = pd.read\_csv('train.csv')  
# Total passenger count  
size = len(df)  
  
# Task 1  
**print** "Task 1. Sex distribution:\n", df['Sex'].value\_counts() / size  
  
# Task 2  
**print** "Task 2. Class distribution\n", df['Pclass'].value\_counts(sort=False) / size  
  
# Task 3&4  
**print** "Task 3&4. Mean age dependency:\n", df.groupby(['Sex', 'Pclass'])['Age'].mean()  
  
# Task 5  
**print** "Task 5. Proportion of survived:", float(len(df[df['Survived'] == 1])) / size  
  
# Task 6  
**print** "Task 6. Mean age:", df['Age'].mean(), "\nMean ages grouped by survived:\n", \  
 df.groupby('Survived')['Age'].mean()  
  
# Task 7  
**print** "Task 7. Proportion of survived by sex:\n", df[df['Survived'] == 1]['Sex'].value\_counts() / \  
 df['Sex'].value\_counts()  
# Task 8  
**print** "Task 8. Mean ticket price:", df['Fare'].mean(), "\nMean absolute deviation of prices:", \  
 df['Fare'].mad()  
  
# Task 9  
**print** "Task 9. Mean ticket price among survived and died:\n", \  
 df.groupby('Survived')['Fare'].mean()  
  
# Task 10  
**print** "Task 10. The most popular male name:", df[df['Sex'] == 'male']['Name'].str.extract(  
 '(Mr.|Master.|Don.|Rev.) (\w+)')[1].value\_counts().idxmax()