**Київський національний університет імені Тараса Шевченка**

**Факультет комп’ютерних наук та кібернетики**

**ЗВІТ**

**Лабораторна робота №2**

|  |
| --- |
| Студента групи ІПС-21 |
| Ольховатого Ігоря Васильовича |

**ЗМІСТ**

[**1.** **Опис вхідної інформації** 3](#_Toc114681025)

[**2.** **Постановка задачі** 4](#_Toc114681026)

[**3.** **Опис додаткового теоретичного матеріалу** 5](#_Toc114681027)

[**4.** **Покроковий опис обробки даних** 6](#_Toc114681028)

[**5.** **Висновки** 6](#_Toc114681029)

[**6.** **Додаток. Програмна реалізація** 15](#_Toc114681030)

[**7.** **Список використаних джерел** 16](#_Toc114681031)

# **Опис вхідної інформації**

Затоплення «Титаніка» — одна з найгірших корабельних аварій в історії.

15 квітня 1912 року під час свого першого плавання корабель RMS Titanic, який вважався «непотоплюваним», затонув після зіткнення з айсбергом. На жаль, не вистачило рятувальних шлюпок для всіх на борту, що призвело до загибелі 1502 з 2224 пасажирів і членів екіпажу.

Оригінальний набір даних взятий з джерела: [*https://www.kaggle.com/competitions/titanic/data?select=train.csv*](https://www.kaggle.com/competitions/titanic/data?select=train.csv)

Відповідні спостереження (із використаних в рамках даної лабораторної роботи) містять інформацію про:

* **Pclass** (клас пасажира, одне із значень від 1 до 3 включно позначає відповідно «перший», «другий» та «третій» клас. У соціально-економічному сенсі «перший» клас вважається «вищим», «другий» клас вважається «середнім», «третій» вважається «нижчим».)
* **Age** (вік пасажира у роках. Вік є дробовим, якщо менше 1. Якщо вік приблизний, чи має він форму xx.5)
* **Fare** (ціна квитка у доларах США на момент покупки)
* **FamilyMembers** (Кількість членів сім’ї на борту. Це сума змінних **SibSp** (кількість *родичів*, *членів подружжя* на борту) та **Parch** (кількість *батьків/дітей* на борту) з оригінального набору даних. Даний набір даних визначає сімейні стосунки таким чином: *родич* = брат, сестра, зведений брат, зведена сестра; *член подружжя* = чоловік, дружина (коханки та наречені ігнорувалися); *батьки* = мати, батько; *дитина* = дочка, син, падчериця, пасинок)
* **Survival** (Те, чи вижив пасажир. Дана якісна змінна закодована в бінарному вигляді 0 – пасажир не вижив, 1 – пасажир вижив, і розглядається як кількісна)

Використана частина даних містять інформацію про 714 пасажирів.

# **Постановка задачі**

Множиною скалярних змінних, для яких вирішено з'ясувати істотність їх статистичного зв'язку обрано змінні **Pclass***,* **Age***,* **Fare***,* **Survival.**

Потрібно провести кореляційний аналіз обраних для обробки вище описаних скалярних змінних:

● на основі результатів попередньої обробки обраного набору даних визначитися, які характеристики статистичного зв'язку потрібно використати при подальшому їх кореляційному аналізі

● провести аналіз істотності парних статистичних зв'язків для усіх пар скалярних змінних, навівши для них:

* вибіркове значення відповідної парної характеристики статистичного зв'язку,
* максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим,
* впорядковану послідовність усіх пар скалярних змінних у порядку спадання істотності статистичного зв'язку між ними,
* сформулювати висновки по кореляційному аналізу парних статистичних зв'язків для обраного набору скалярних змінних.

● провести аналіз істотності множинних статистичних зв'язків між кожною обраною в якості залежної скалярною змінною та множиною усіх інших скалярних змінних (які виступають у ролі незалежних змінних), навівши для них:

* вибіркове значення відповідної множинної характеристики статистичного зв'язку,
* максимальний рівень значущості при якому відповідний множинний статистичний зв'язок не є значимим,
* впорядковану послідовність усіх скалярних змінних у порядку спадання істотності множинного статистичного зв'язку їх з множиною усіх інших скалярних змінних,
* сформулювати висновки по кореляційному аналізу множинних статистичних зв'язків для обраного набору скалярних змінних.

# **Опис додаткового теоретичного матеріалу**

Коефіцієнт кореляції Пірсона:

<https://en.wikipedia.org/wiki/Pearson_correlation_coefficient>

Коефіцієнт кореляції Спірмана:

<https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient>

Коефіцієнт кореляції Кендала:

<https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient>

Коефіцієнти кореляції:

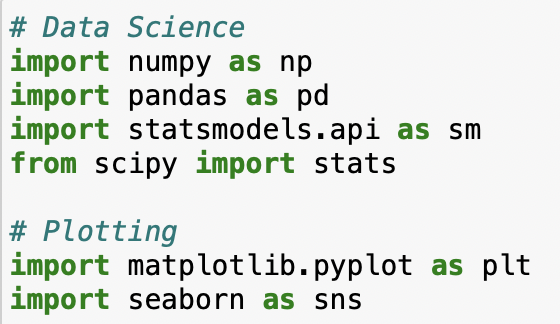
<https://datascience.stackexchange.com/questions/64260/pearson-vs-spearman-vs-kendall>

Тест нормальності Шапіро-Уілка.

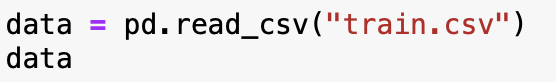
<https://en.wikipedia.org/wiki/Shapiro–Wilk_test>

# **Покроковий опис обробки даних**

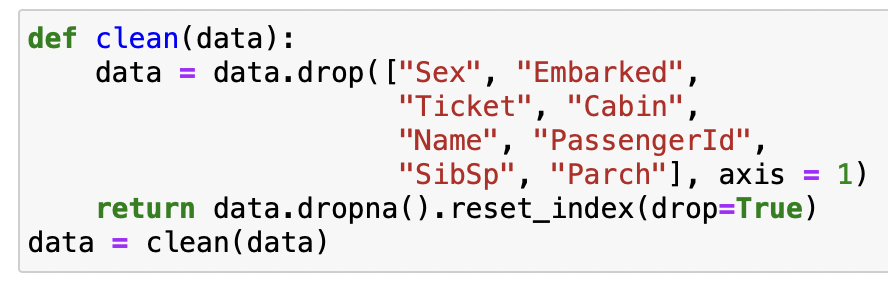
Для початку, імпортуємо відповідні бібліотеки для подальшої роботи.

**

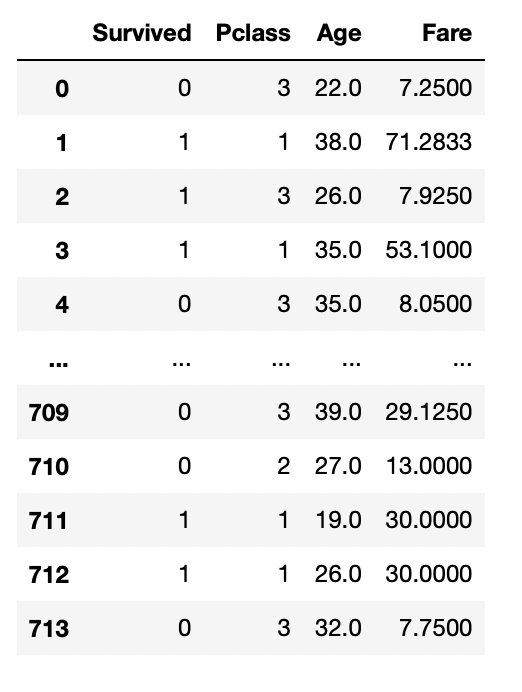
Тепер завантажимо таблицю з потрібним набором даних.

**

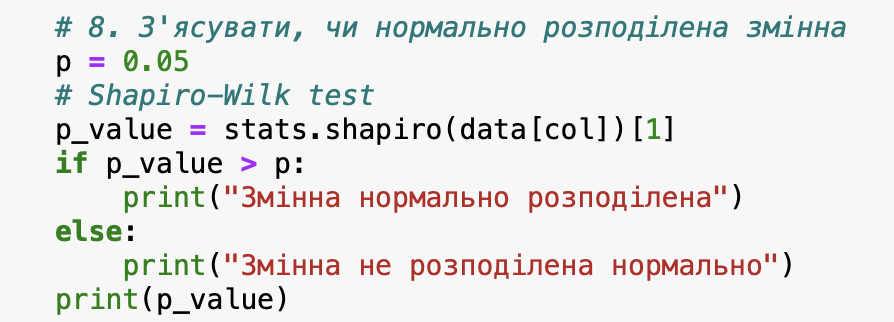
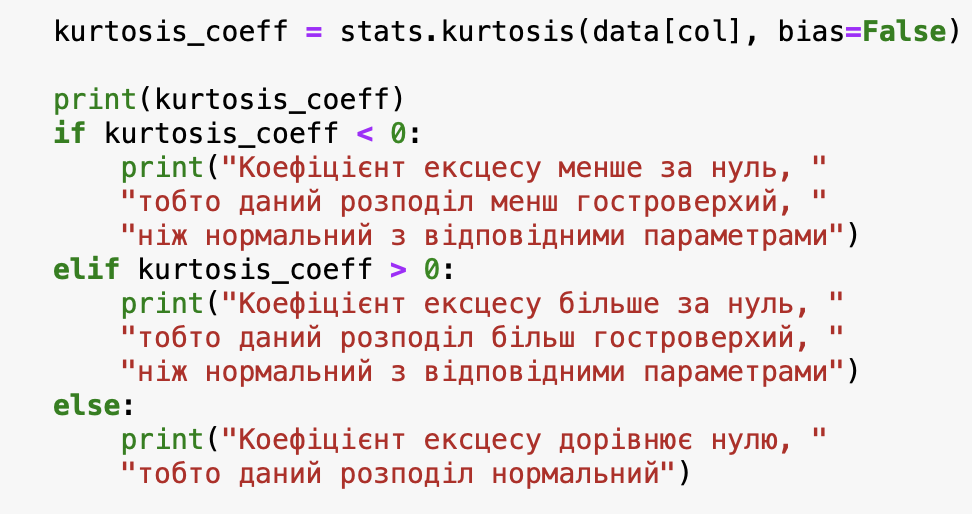
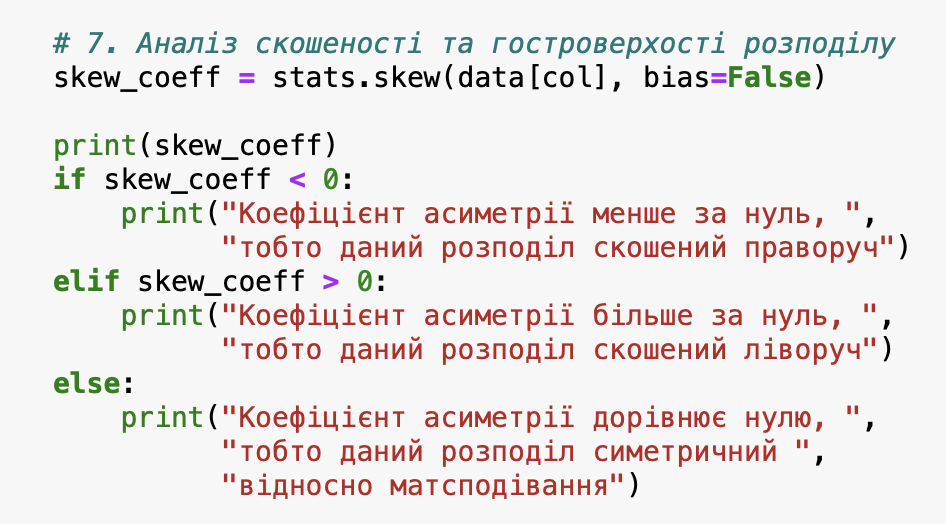
Використовуючи інформацію з попереднього аналізу даних, залишимо лише змінні, щоб будемо аналізувати.



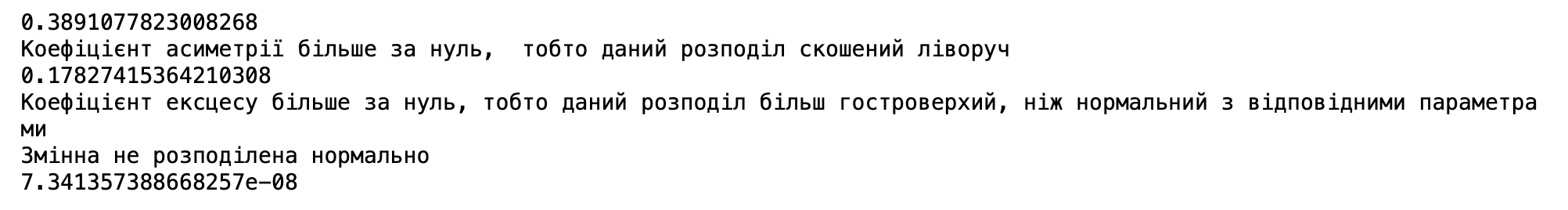
Тепер вхідні дані мають наступний вигляд:



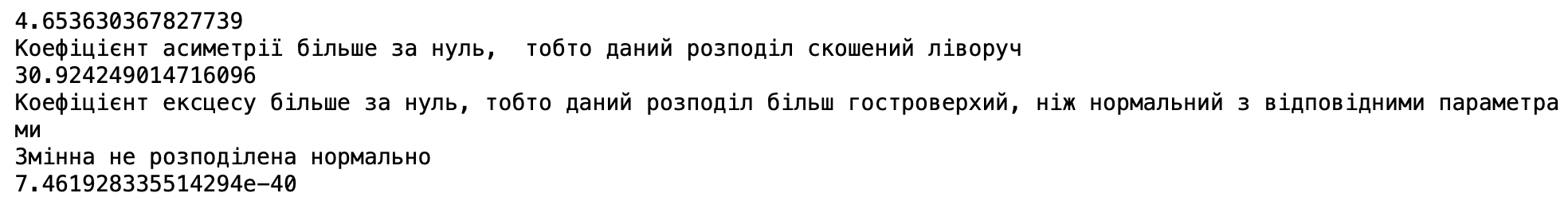
**Перевірка змінних на нормальний розподіл**



Age



Fare



Бачимо, що серед даних змінних немає нормально розподілених. Зауважимо, що для змінних Pclass та survived перевірка на нормальність немає сенсу.

**Аналіз істотності парних статистичних зв'язків**

Оскільки в нашому випадку немає нормально розподілених змінних, а отже не можна використовувати коефіцієнт Пірсона, який робить припущення про нормальний розподіл змінних. Має сенс використовувати непараметричні рангові коефіцієнти Спірмена і Кендела, значення змінних для яких можуть бути неперервними або порядковими, і можуть не мати наближеного до нормального розподілу.

Коефіцієнт Спірмена:

Коефіцієнт Кендела:

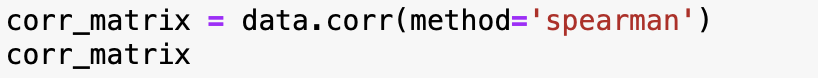
Коефіцієнти можуть набувають значень з інтервалу [-1; 1]. Конкретно в рамках даної лабораторної роботи автор трактує значення коефіцієнту кореляції як:

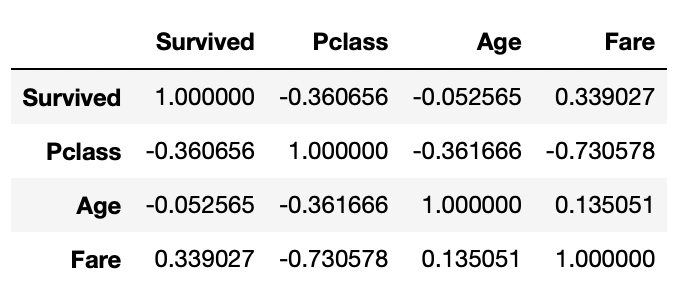
– кореляція відсутня

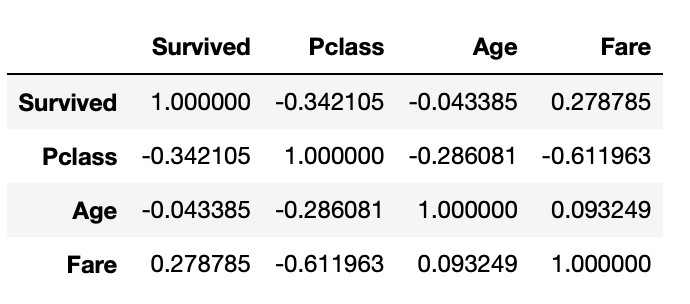
– кореляція низька

– кореляція середня

– кореляція висока







Враховуємо дані з обох коефіцієнтів кореляції для більш об’єктивного дослідження. У якості максимального рівня значущості, при якому відповідний парний статистичний зв'язок не є значимим беремо максимальне значення з отриманих p-value.

**Кореляція кількісних змінних:**

Найбільший за модулем коефіцієнт кореляції у Pclass та Fare. Він високий, але від’ємний, з чого можна зробити висновок, що невеликі збільшення значення номеру класу сильно відповідає зниженню значення ціни квитка. Це логічно, адже найдорожчі квитки на 1 клас, а чим більше номер класу, тим менше вони коштують.

Максимальний рівень значущості, при якому відповідний парний статистичний зв'язок не є значимим (p-value) - 1.96e-97

Середній рівень кореляції між Pclass та Age( невеликі збільшення значення номеру класу відповідає зниженню значення віку). Це підтверджує наше попереднє припущення (з лабораторної роботи №1 «Попередній аналіз даних»), що першим класом подорожували люди більш поважного віку, а другим та третім у більшості молодь.

Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) - 1.72e-22

Низький коефіцієнт кореляції у Age та Fare. Він додатній, тобто люди старшого віку зазвичай мали дорожчі квитки. Це доповнює нашу гіпотезу, що старші люди подорожували першим дорожчим класом, але варто зазначити, що коефіцієнт кореляції досить низький, тому робити такий висновок на основі лише його не було б коректно (у даному випадку він радше точно не суперечить нашій гіпотезі).

Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) - 0.00029

**Кореляція з категоріальною змінною Survived:**

(яка для зручності роботи з нею, у тому числі підрахунку відповідних коефіцієнтів кореляції представлена у бінарному вигляді, як було описано раніше)

Кореляція із змінною Age настільки низька, що нею можна знехтувати, з чого робимо висновок що вік не впливав на те, чи виживе пасажир.

Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) – 0.16

Кореляція із змінною Fare свідчить про те, що чим люди які платили більше інколи мали більшу імовірність вижити.

Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) – 1.39e-19

Кореляція із змінною Pclass підтверджує попередню гіпотезу, адже більший номер класу означає дешевший білет.

Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) – 5.95e-22

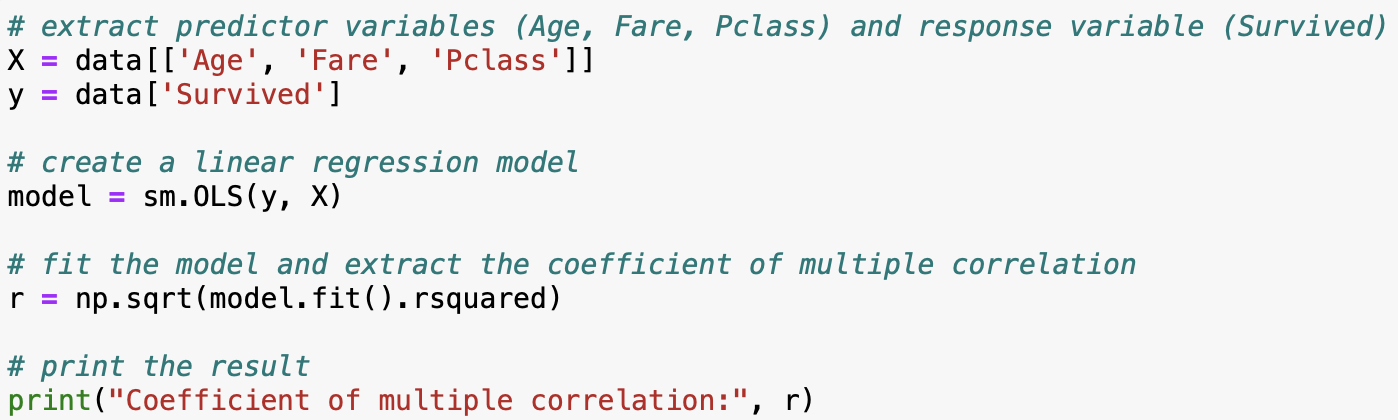
Упорядкована послідовність усіх пар скалярних змінних у порядку спадання істотності статистичного зв'язку між ними:

1. *(Pclass, Fare)*
2. *(Survived, Pclass)*
3. *(Pclass, Age)*
4. *(Survived, Fare)*
5. *(Fare, Age)*
6. *(Survived, Age)*

**Аналіз істотності множинних статистичних зв'язків**

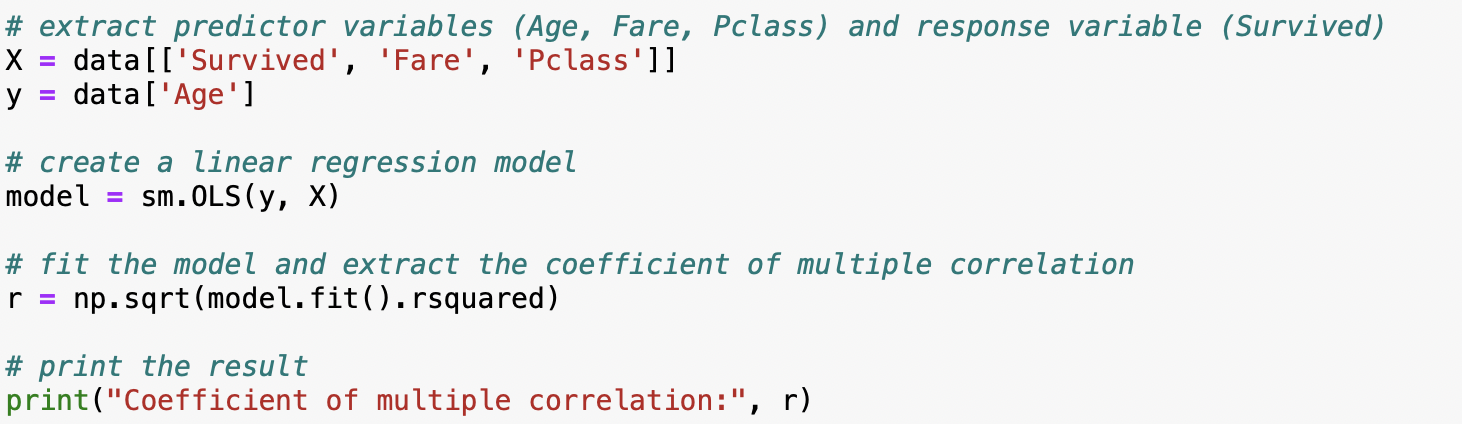
Для аналізу множинних зв’язків будемо використовувати так званий множинний коефіцієнт детермінації.

, де Для цього створимо модель лінійної регресії для кожної змінної, яка за допомогою звичайного методу найменших квадратів натренується на нашому наборі даних. Далі ми легко зможемо отримати вибіркове значення множинної характеристики статистичного зв'язку.

**

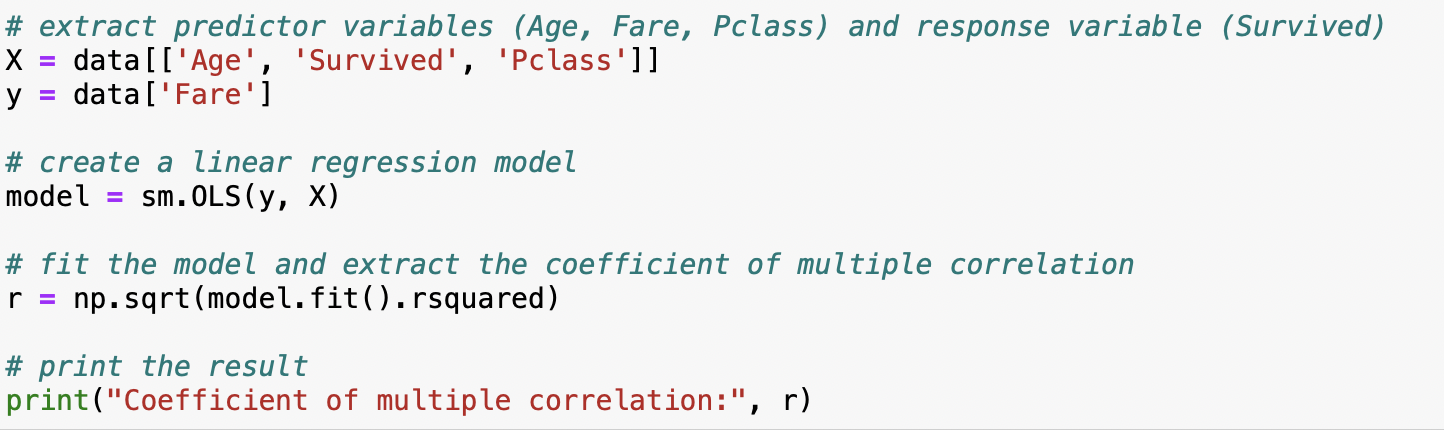
Множинний коефіцієнт детермінації = 0.6296466518755806.

Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) – 0.00084



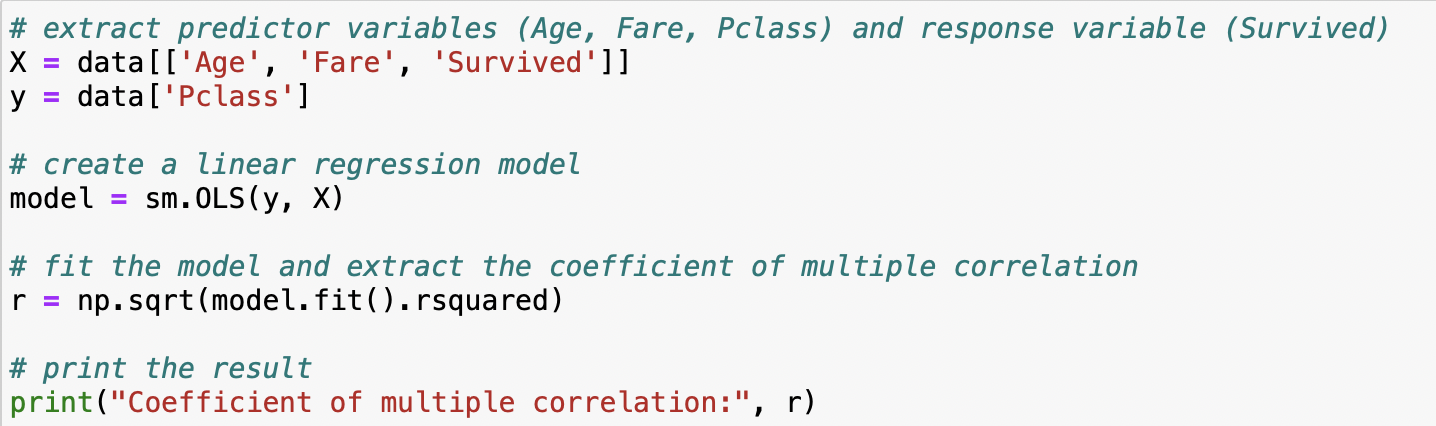
Множинний коефіцієнт детермінації = 0.8324992102290122.

Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) – 0.00084



Множинний коефіцієнт детермінації = 0.6138236093319311.

Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) – 2.24-e6



Множинний коефіцієнт детермінації = 0.7964297781062655.

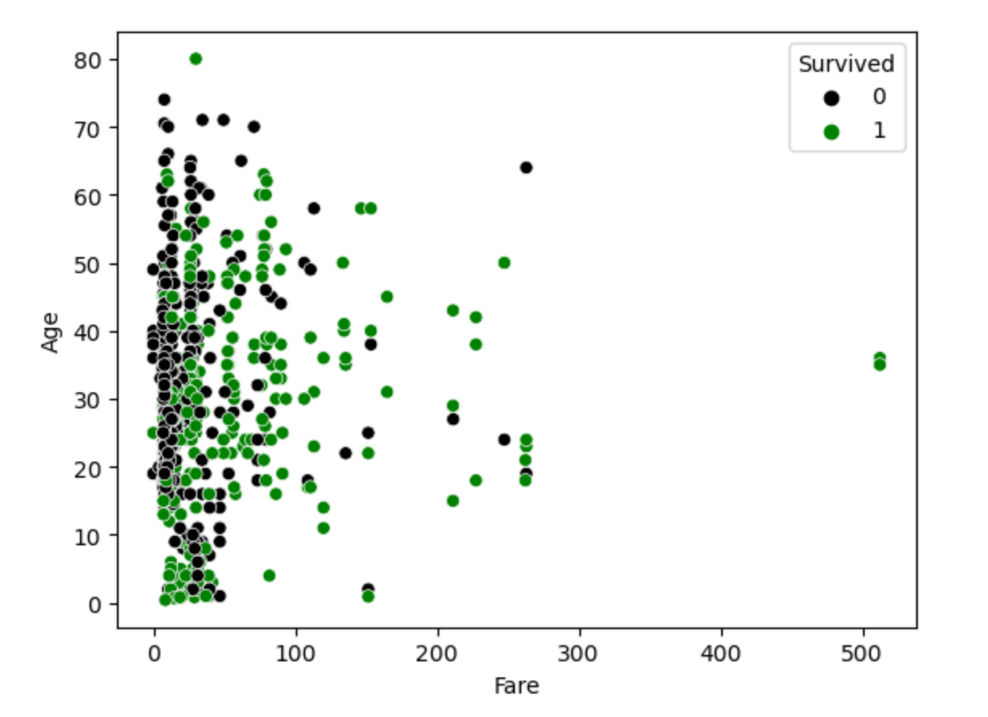
Максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим (p-value) – 2.24-e6

Упорядкована послідовність усіх скалярних змінних у порядку спадання істотності множинного  статистичного зв'язку їх з множиною усіх інших скалярних змінних:

1. *Age, (Survived, Fare, Pclass)*
2. *Pclass, (Survived, Fare, Age)*
3. *Survived, (Age, Fare, Pclass)*
4. *Fare, (Survived, Age, Pclass)*

Дані про множинний коефіцієнт детермінації можна трактувати як те, яка частина дисперсії залежної змінної визначається варіацією функції регресії. Можна зробити висновок, що змінні досить сильно пов’язані між собою. Досить цікаво, що найбільший коефіцієнт саме для Age та (Survived, Fare, Pclass).

На останок, пропонується подивитися на ілюстрацію кореляції між Age та Fare у залежності від Survived.



# **Висновки**

Я провів кореляційний аналіз для кожної із для усіх пар скалярних змінних, навівши для них вибіркове значення відповідної парної характеристики статистичного зв'язку, максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим, впорядковану послідовність усіх пар скалярних змінних у порядку спадання істотності статистичного зв'язку між ними. Окрім того, проведено аналіз істотності множинних статистичних зв'язків між кожною обраною в якості залежної скалярною змінною та множиною усіх інших скалярних змінних та наведено відповідні характеристики.

Аналіз даного набору даних показав, що виживання пасажиру дійсно залежало від інших змінних. Аналіз підтвердив, що пасажири похилого віку дійсно у більшості подорожували 1 класом і мали вищий економічний статус. Цікаво, що вік пасажирів майже не впливав на те, вижили вони під час катастрофи чи ні, проте ціна білету мала більший вплив.

Дані, отримані під час кореляційного аналізу, можуть бути використані в подальшому для побудови моделі, що може вивчати зв’язки між змінними з даного набору даних.

Робота з мовою програмування “Python” і відповідними програмними пакетами була зручною та простою. До плюсів можна віднести зрозумілу документацію, ефективну роботу та багатий вибір засобів для ілюстрацій.

# **Додаток. Програмна реалізація**

# Data Science

import numpy as np

import pandas as pd

import statsmodels.api as sm

from scipy import stats

# Plotting

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read\_csv("train.csv")

def clean(data):

data = data.drop(["Sex", "Embarked",

"Ticket", "Cabin",

"Name", "PassengerId",

"SibSp", "Parch"], axis = 1)

return data.dropna().reset\_index(drop=True)

data = clean(data)

data

for col in ['Fare', 'Age']:

print(col)

# 7. Аналіз скошеності та гостроверхості розподілу

skew\_coeff = stats.skew(data[col], bias=False)

print(skew\_coeff)

if skew\_coeff < 0:

print("Коефіцієнт асиметрії менше за нуль, ",

"тобто даний розподіл скошений праворуч")

elif skew\_coeff > 0:

print("Коефіцієнт асиметрії більше за нуль, ",

"тобто даний розподіл скошений ліворуч")

else:

print("Коефіцієнт асиметрії дорівнює нулю, ",

"тобто даний розподіл симетричний ",

"відносно матсподівання")

kurtosis\_coeff = stats.kurtosis(data[col], bias=False)

print(kurtosis\_coeff)

if kurtosis\_coeff < 0:

print("Коефіцієнт ексцесу менше за нуль, "

"тобто даний розподіл менш гостроверхий, "

"ніж нормальний з відповідними параметрами")

elif kurtosis\_coeff > 0:

print("Коефіцієнт ексцесу більше за нуль, "

"тобто даний розподіл більш гостроверхий, "

"ніж нормальний з відповідними параметрами")

else:

print("Коефіцієнт ексцесу дорівнює нулю, "

"тобто даний розподіл нормальний")

# 8. З'ясувати, чи нормально розподілена змінна

p = 0.05

# Shapiro-Wilk test

p\_value = stats.shapiro(data[col])[1]

if p\_value > p:

print("Змінна нормально розподілена")

else:

print("Змінна не розподілена нормально")

print(p\_value)

corr\_matrix = data.corr(method='spearman')

corr\_matrix = data.corr(method='kendall') stats.spearmanr(data['Survived'], data['Age']).pvalue stats.kendalltau(data['Survived'], data['Age']).pvalue

# extract predictor variables (Age, Fare, Pclass) and response variable (Survived)

X = data[['Age', 'Fare', 'Survived']]

y = data['Pclass']

# create a linear regression model

model = sm.OLS(y, X)

# fit the model and extract the coefficient of multiple correlation

r = np.sqrt(model.fit().rsquared)

# print the result

print("Coefficient of multiple correlation:", r)

model.fit().pvalues.max()

sns.scatterplot(data=data, x='Fare', y='Age', hue='Survived', palette=['black', 'green'], )

**Список використаних джерел**

1. Документація мови програмування Python

<https://www.python.org/doc/>

1. Документація бібліотеки Pandas для роботи з різними структурами даних, у тому числі з таблицями.

<https://pandas.pydata.org/docs/>

1. Документація бібліотеки NumPy для роботи з науковими обчисленнями.

<https://numpy.org/doc/>

1. Документація бібліотеки SciPy для роботи з математичними функціями.

<https://docs.scipy.org/doc//scipy/index.html>

1. Документація бібліотеки Seaborn для роботи з графіками.

<https://seaborn.pydata.org>

1. Документація бібліотеки MatPlotLib для роботи з графіками, на базі якої працює Seaborn.

<https://matplotlib.org>

1. Деталі підрахунку коефіцієнту варіації.

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.variation.html>

1. Деталі підрахунку коефіцієнту скошеності.

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html>

1. Деталі підрахунку коефіцієнту ексцесу (за Фішером).

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kurtosis.html>