# Лабораторна робота № 2. Аналіз та візуалізація великих наборів даних

**Мета:** ознайомити здобувачів з основами аналізу даних та візуалізації з використанням бібліотек Pandas та Seaborn в Python. Навчити здобувачів завантажувати дані з CSV-файлів, проводити попередню обробку даних, виконувати статистичний аналіз та будувати інформативні візуалізації. Розвинути навички командної роботи, взаємодії та послідовного виконання завдань.

## Теоретичні відомості

Великі набори даних (Big Data) відіграють ключову роль у сучасному світі, охоплюючи різні сфери діяльності, від науки та бізнесу до державного управління та соціальних досліджень. Ефективний аналіз та візуалізація таких даних є важливими інструментами для отримання цінних інсайтів, виявлення прихованих закономірностей та підтримки прийняття обґрунтованих рішень.

Pandas – це потужна бібліотека Python, яка надає високоефективні, прості у використанні структури даних та інструменти аналізу даних. Вона спеціально розроблена для роботи з табличними даними, такими як дані з електронних таблиць, баз даних або CSV-файлів. Основними структурами даних в Pandas є Series (одномірний масив з мітками) та DataFrame (двовимірна таблиця з мітками рядків та стовпців). Pandas надає широкий спектр функцій для завантаження, очищення, перетворення, фільтрації, агрегації та аналізу даних.

Seaborn – це бібліотека Python для створення статистичних графіків, яка базується на бібліотеці Matplotlib. Seaborn надає високорівневий інтерфейс для створення інформативних та естетично привабливих візуалізацій даних. Вона тісно інтегрована з Pandas, що дозволяє легко будувати графіки на основі DataFrame. Seaborn автоматично обробляє багато деталей, пов'язаних з візуалізацією, таких як вибір кольорової палітри, налаштування осей та легенд, що дозволяє зосередитися на інтерпретації даних, а не на технічних деталях.

# Приклад виконання роботи

## Завдання

1. Завантажити набір даних «recent-grads.csv» у DataFrame Pandas. Цей набір даних містить інформацію про випускників коледжів США за 2010-2012 роки, включаючи демографічні дані, спеціальності, рівень зайнятості та доходи.
2. Провести попередню обробку даних: видалити стовпці, які не потрібні для аналізу, та обробити пропущені значення.
3. **Здобувач 1**. Розрахувати середній рівень безробіття серед випускників.
4. **Здобувач 1**. Побудувати гістограму розподілу медіанного заробітку випускників.
5. **Здобувач 1**. Знайти медіану середнього заробітку випускників, які мають частку жінок у спеціальності більше 0.5.
6. **Здобувач 2**. Використовуючи дані, передані **Здобувачем 1**, побудувати діаграму розсіювання, що показує залежність між часткою жінок у спеціальності та медіанним заробітком. Додати на графік горизонтальну лінію, що відповідає медіані середнього заробітку випускників, які мають частку жінок у спеціальності більше 0.5, розраховану **Здобувачем 1**.
7. **Здобувач 2**. Використовуючи дані, передані **Здобувачем 1**, побудувати ящик з вусами для медіанного заробітку випускників різних категорій спеціальностей.
8. **Здобувач 2**. Розрахувати **secret key** як суму індексів (номер рядка) шести спеціальностей з найвищим рівнем безробіття, використовуючи дані, передані **Здобувачем 1**.

**Контрольне запитання**

Як співвідноситься частка жінок у спеціальності з медіанним заробітком випускників? Чи є суттєва різниця у медіанному заробітку між різними категоріями спеціальностей?

Відповідь: _____

## Хід роботи

0. Комунікація та планування (виконується обома учасниками).

Обговоріть завдання лабораторної роботи та переконайтеся, що обидва учасники розуміють цілі, завдання та послідовність виконання.

Сплануйте етапи виконання роботи, враховуючи дедлайни та час, необхідний для кожного етапу. Домовтесь про спосіб передачі даних між учасниками (наприклад, збереження проміжних результатів у CSV-файл, або безпосередньо в Google Colab).

1. Імпорт бібліотек та налаштування середовища (виконує **Здобувач 1**).

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

sns.set_style("whitegrid")
```

На цьому етапі ми імпортуємо необхідні бібліотеки: Pandas для роботи з даними та Seaborn для візуалізації. Ми також налаштовуємо стиль графіків Seaborn, щоб зробити їх більш привабливими.

2. Завантаження даних (виконує **Здобувач 1**).

```
df = pd.read_csv("recent-grads.csv")
df.head(10)
```

| | Rank | Major_code | Major | Total | Men | Women | Major_category | ShareWomen | Sample_size | Employed | ... | Part_time | Full_t: |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2419 | PETROLEUM ENGINEERING | 2339.0 | 2057.0 | 282.0 | Engineering | 0.120564 | 36 | 1976 | ... | 270 | |
| **1** | 2 | 2416 | MINING AND MINERAL ENGINEERING | 756.0 | 679.0 | 77.0 | Engineering | 0.101852 | 7 | 640 | ... | 170 | |
| **2** | 3 | 2415 | METALLURGICAL ENGINEERING | 856.0 | 725.0 | 131.0 | Engineering | 0.153037 | 3 | 648 | ... | 133 | |
| **3** | 4 | 2417 | NAVAL ARCHITECTURE AND MARINE ENGINEERING | 1258.0 | 1123.0 | 135.0 | Engineering | 0.107313 | 16 | 758 | ... | 150 | |
| **4** | 5 | 2405 | CHEMICAL ENGINEERING | 32260.0 | 21239.0 | 11021.0 | Engineering | 0.341631 | 289 | 25694 | ... | 5180 | |
| **5** | 6 | 2418 | NUCLEAR ENGINEERING | 2573.0 | 2200.0 | 373.0 | Engineering | 0.144967 | 17 | 1857 | ... | 264 | |
| **6** | 7 | 6202 | ACTUARIAL SCIENCE | 3777.0 | 2110.0 | 1667.0 | Business | 0.441356 | 51 | 2912 | ... | 296 | |
| **7** | 8 | 5001 | ASTRONOMY AND ASTROPHYSICS | 1792.0 | 832.0 | 960.0 | Physical Sciences | 0.535714 | 10 | 1526 | ... | 553 | |
| **8** | 9 | 2414 | MECHANICAL ENGINEERING | 91227.0 | 80320.0 | 10907.0 | Engineering | 0.119559 | 1029 | 76442 | ... | 13101 | |
| **9** | 10 | 2408 | ELECTRICAL ENGINEERING | 81527.0 | 65511.0 | 16016.0 | Engineering | 0.196450 | 631 | 61928 | ... | 12695 | |

10 rows × 21 columns

Рисунок 1.1. Результат виконання завдання №2

Тут ми використовуємо функцію `read_csv()` з бібліотеки Pandas, щоб завантажити дані з файлу «recent-grads.csv» у DataFrame, який ми назвали `df`. Цей файл повинен бути заздалегідь завантажений у робоче середовище Google Colab.

3. Попередня обробка даних (виконує **Здобувач 1**).

```
df = df.drop(['Rank', 'P25th', 'P75th'], axis=1)

df = df.dropna()
df.head(10)
```

| | Major_code | Major | Total | Men | Women | Major_category | ShareWomen | Sample_size | Employed | Full_time | Part_time | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2419 | PETROLEUM ENGINEERING | 2339.0 | 2057.0 | 282.0 | Engineering | 0.120564 | 36 | 1976 | 1849 | 270 | |
| 1 | 2416 | MINING AND MINERAL ENGINEERING | 756.0 | 679.0 | 77.0 | Engineering | 0.101852 | 7 | 640 | 556 | 170 | |
| 2 | 2415 | METALLURGICAL ENGINEERING | 856.0 | 725.0 | 131.0 | Engineering | 0.153037 | 3 | 648 | 558 | 133 | |
| 3 | 2417 | NAVAL ARCHITECTURE AND MARINE ENGINEERING | 1258.0 | 1123.0 | 135.0 | Engineering | 0.107313 | 16 | 758 | 1069 | 150 | |
| 4 | 2405 | CHEMICAL ENGINEERING | 32260.0 | 21239.0 | 11021.0 | Engineering | 0.341631 | 289 | 25694 | 23170 | 5180 | |
| 5 | 2418 | NUCLEAR ENGINEERING | 2573.0 | 2200.0 | 373.0 | Engineering | 0.144967 | 17 | 1857 | 2038 | 264 | |
| 6 | 6202 | ACTUARIAL SCIENCE | 3777.0 | 2110.0 | 1667.0 | Business | 0.441356 | 51 | 2912 | 2924 | 296 | |

Рисунок 1.2. Результат виконання завдання №3

На цьому етапі ми видаляємо стовпці «Rank», «P25th» та «P75th», які не будуть використовуватися в нашому аналізі, використовуючи метод `drop()`. Потім ми видаляємо рядки, які містять пропущені значення, за допомогою методу `dropna()`. Це важливо зробити, оскільки багато функцій аналізу та візуалізації не можуть коректно працювати з даними, що містять пропуски.

4. Розрахунок середнього рівня безробіття (виконує **Здобувач 1**).

```
df_variant35 = df[df['Major_category'].isin(["Law & Public Policy",
"Communications & Journalism"])]

employment_rate = df_variant35.groupby('Major')['Employed'].sum() /
(df_variant35.groupby('Major')['Employed'].sum() +
df_variant35.groupby('Major')['Unemployed'].sum())
print("Середній рівень працевлаштування для кожної спеціальності:")
print(employment_rate)

average_employment_rate_variant = employment_rate.mean()
print(f"Середній рівень працевлаштування для варіанту:
{average_employment_rate_variant}")

employment_rate_sorted = employment_rate.sort_values()
```

```
Середній рівень працевлаштування для кожної спеціальності:
Major
ADVERTISING AND PUBLIC RELATIONS        0.932039
COMMUNICATIONS                          0.924823
COURT REPORTING                         0.988310
CRIMINAL JUSTICE AND FIRE PROTECTION    0.917548
JOURNALISM                              0.930824
MASS MEDIA                              0.910163
PRE-LAW AND LEGAL STUDIES               0.928035
PUBLIC ADMINISTRATION                   0.840509
PUBLIC POLICY                           0.871574
dtype: float64
Середній рівень працевлаштування для варіанту: 0.915980575920436
```

Рисунок 1.3. Результат виконання завдання №4

Тут ми групуємо дані за спеціальностями «Major» за допомогою методу `groupby()`, а потім розраховуємо середній рівень безробіття «Unemployment_rate» для кожної спеціальності за допомогою методу `mean()`. Результат зберігається у змінній `average_unemployment_rate` та виводиться на екран.

5. Побудова гістограми розподілу медіанного заробітку (виконує **Здобувач 1**).

```
plt.figure(figsize=(12, 6))
sns.histplot(df_variant35['Total'], bins=20, kde=True)
plt.title('Розподіл кількості студентів на спеціальностях категорій Law & Public Policy та Communications & Journalism')
plt.xlabel('Кількість студентів')
plt.ylabel('Кількість спеціальностей')
plt.show()
```

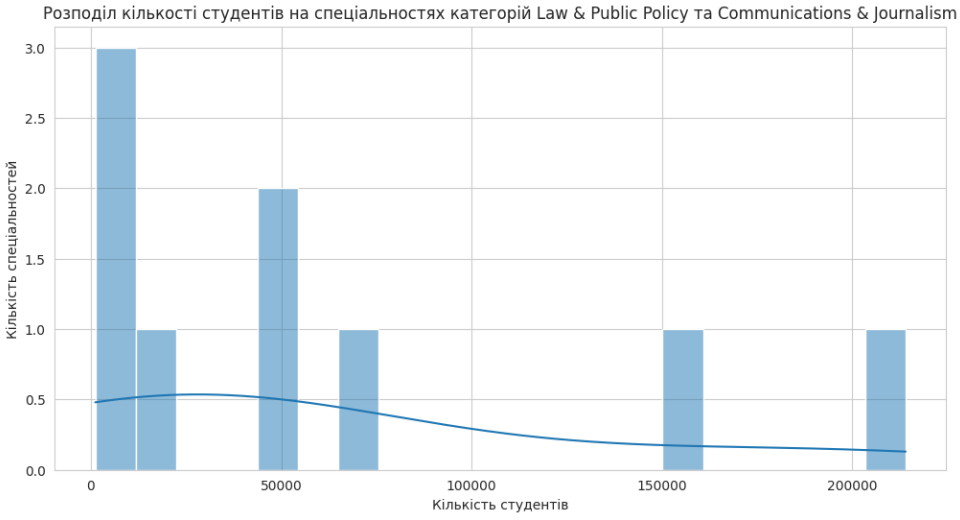


Рисунок 1.4. Результат виконання завдання №5

Ми використовуємо функцію `histplot()` з бібліотеки Seaborn для побудови гістограми розподілу медіанного заробітку «Median». Параметр `bins` задає кількість стовпчиків гістограми, а `kde=True` додає криву оцінки щільності розподілу. Ми також додаємо заголовок та підписи осей для кращого розуміння графіку.

6. Розрахунок медіани середнього заробітку для спеціальностей з часткою жінок > 0.5 (виконує **Здобувач 1**).

```
df_filtered = df_variant35[df_variant35.apply(lambda x: (x['Employed'] / (x['Employed'] + x['Unemployed'])) > average_employment_rate_variant, axis=1)]

median_men_high_employment_rate = df_filtered['Men'].median()

print(f"Медіана кількості чоловіків для спеціальностей з рівнем працевлаштування вище середнього: {median_men_high_employment_rate}")
```

```
Медіана кількості чоловіків для спеціальностей з рівнем працевлаштування вище середнього: 18299.0
```

Рисунок 1.5. Результат виконання завдання №6

7. Передача даних **Здобувачу 2** (виконує **Здобувач 1**).

```
employment_rate_sorted.to_csv("employment_rate_variant35.csv")
df_variant35.to_csv("variant35_data.csv", index=False)

print(f"Передайте Учаснику 2 наступне значення: {median_men_high_employment_rate}")
```

```
Передайте Учаснику 2 наступне значення: 18299.0
```

Рисунок 1.6. Результат виконання завдання №7

8. Отримання даних від **Здобувача 1** (виконує **Здобувач 2**).

```
employment_rate_sorted = pd.read_csv("employment_rate_variant35.csv")
df_variant35 = pd.read_csv("variant35_data.csv")

median_men_high_employment_rate = 18299.0 # Замініть на значення, передане Учасником 1!
```

9. Побудова діаграми розсіювання з горизонтальною лінією (виконує **Здобувач 2**).

```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Men', y='Women', data=df_variant35)
plt.title('Залежність кількості жінок від кількості чоловіків на
спеціальностях (Варіант 35)')
plt.xlabel('Кількість чоловіків')
plt.ylabel('Кількість жінок')
plt.axhline(y=median_men_high_employment_rate, color='r', linestyle='--',
label=f'Медіана чоловіків: {median_men_high_employment_rate}')
plt.legend()
plt.show()
```

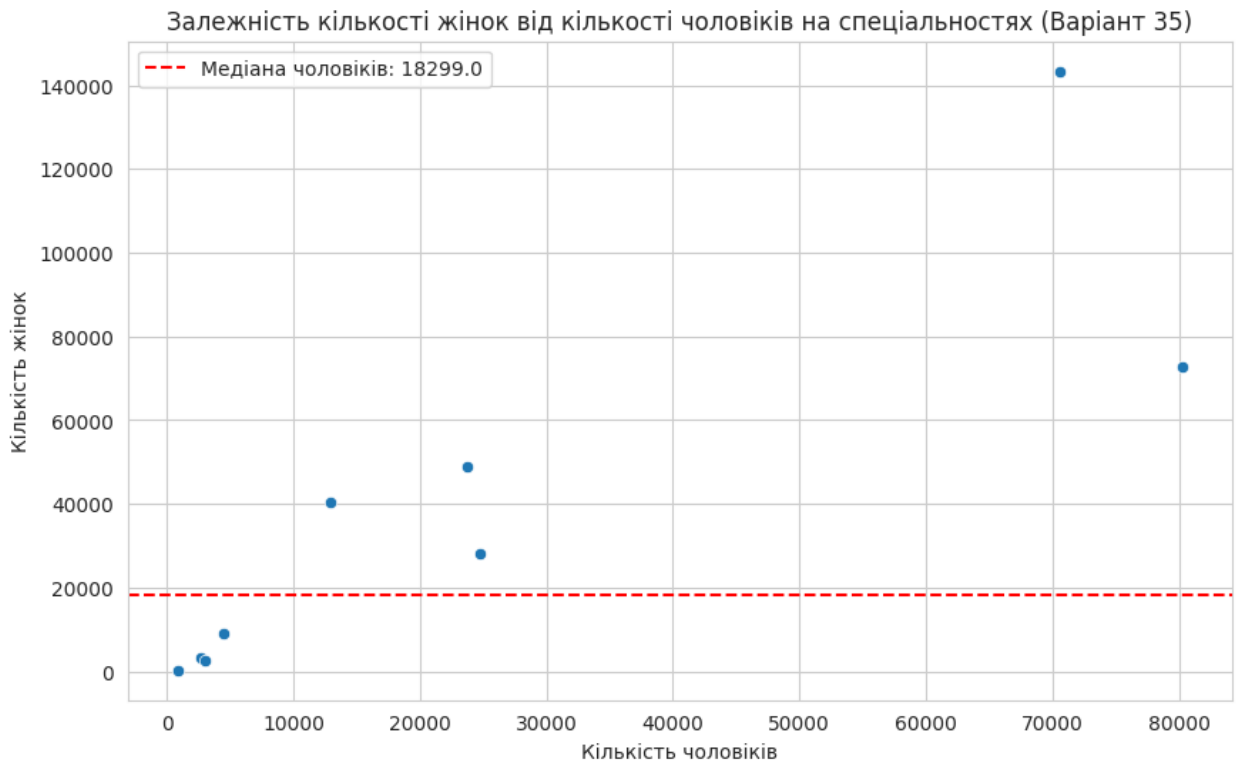


Рисунок 1.7. Результат виконання завдання №9

Ми використовуємо функцію `scatterplot()` з Seaborn для побудови діаграми розсіювання, яка показує залежність між часткою жінок у спеціальності «ShareWomen» та медіанним заробітком «Median». Кожна точка на графіку відповідає одній спеціальності.

10. Побудова ящика з вусами (виконує **Здобувач 2**).

```
plt.figure(figsize=(12, 8))
sns.boxplot(x='Major_category', y='Full_time', data=df_variant35)
plt.title('Розподіл кількості повного робочого дня для спеціальностей
(Варіант 35)')
plt.xlabel('Категорія спеціальності')
plt.ylabel('Кількість повного робочого дня')
plt.xticks(rotation=90)
plt.show()
```

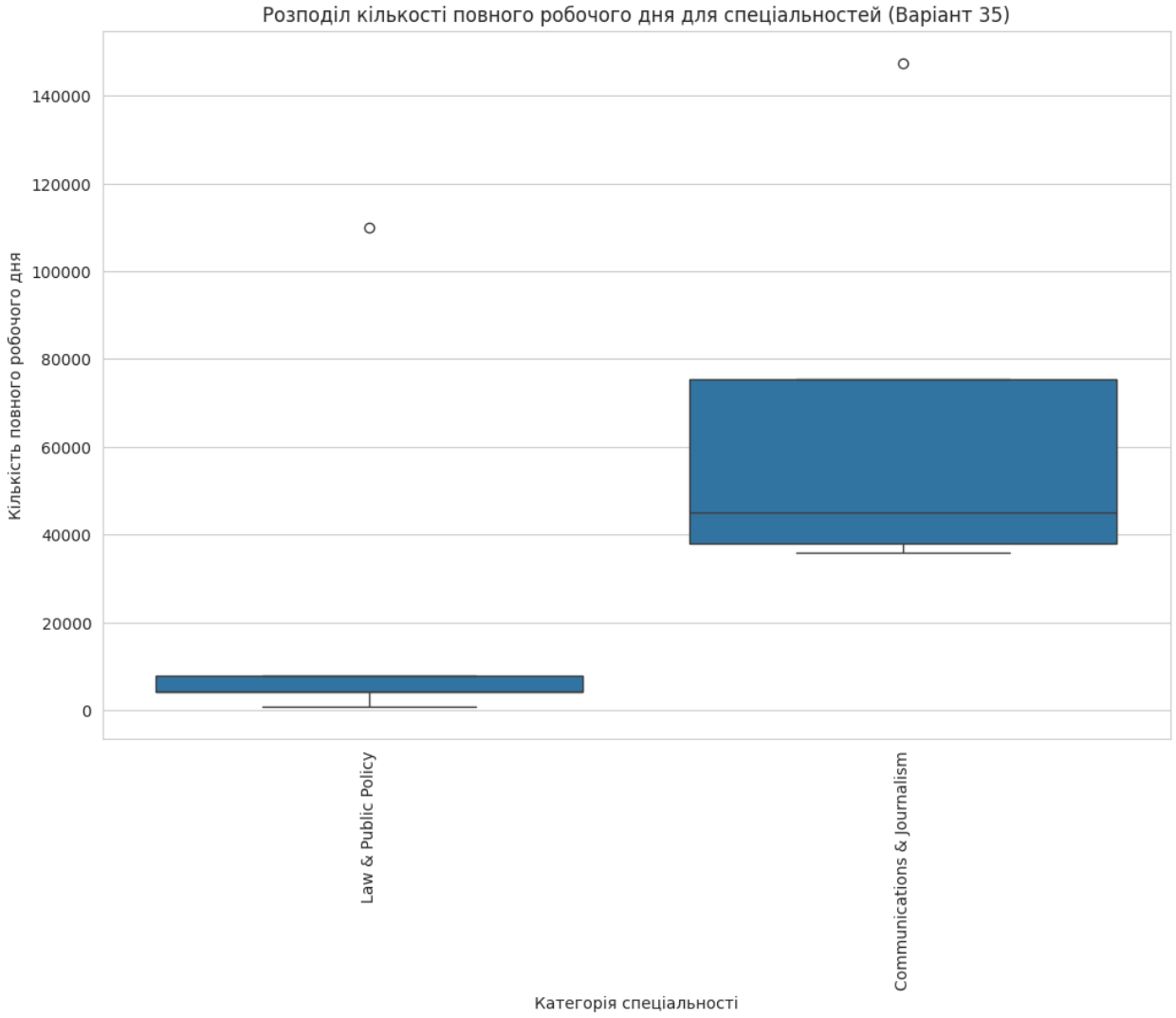


Рисунок 1.8. Результат виконання завдання №10

Ми використовуємо функцію `boxplot()` з Seaborn для побудови ящика з вусами, який показує розподіл медіанного заробітку для різних категорій спеціальностей «Major_category». Цей тип графіку дозволяє порівняти медіану, квартилі та викиди для

різних груп даних. Ми також повертаємо підписи осі x на 90 градусів для кращої читабельності.

11. Розрахунок **secret key** (виконує **Здобувач 2**).

```
employment_rate_sorted = employment_rate.sort_values()
last_6_majors = employment_rate_sorted.head(6).index.tolist()
last_6_indices =
df_variant35.loc[df_variant35['Major'].isin(last_6_majors)].index.tolist()
unique_indices = []
[unique_indices.append(x) for x in last_6_indices if x not in
unique_indices]
checked_indices = [idx if idx != 0 else 1 for idx in unique_indices]


# Розраховуємо secret key як суму індексів
from math import prod

secret_key = prod(checked_indices)

print(f"Secret key: {secret_key}")
```

**Secret key:** 360360

12. Обговорення результатів та відповідь на контрольне запитання (виконують обидва здобувачі).

1. Обговоріть отримані результати та зробіть висновки на основі проведеного аналізу та візуалізацій.
2. Підготуйте спільну відповідь на контрольне запитання.
3. Перевірте secret key на https://keychecker.vercel.app/.

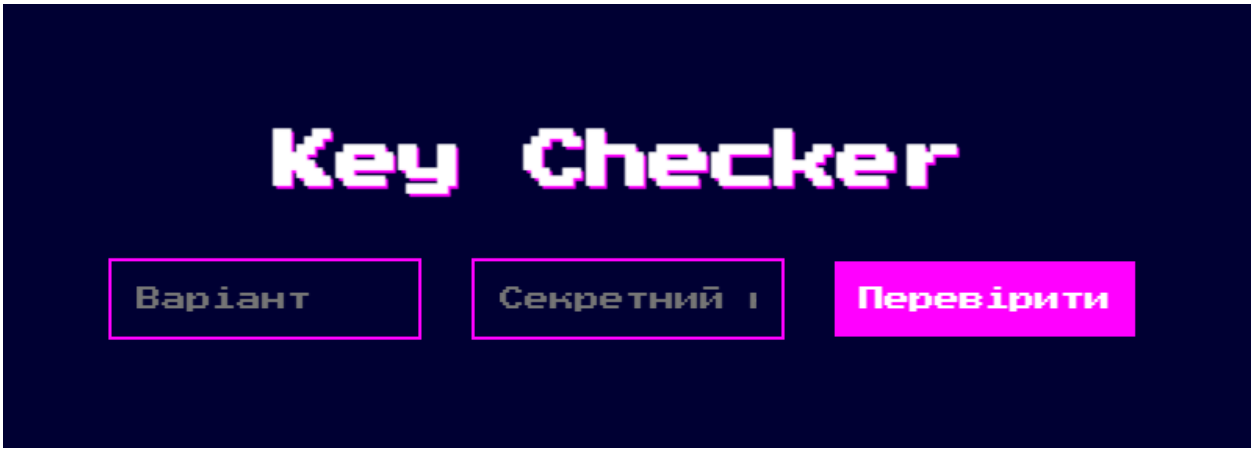


Рисунок 1.9. Результат виконання завдання №11

**Контрольне запитання**

Як співвідноситься частка жінок у спеціальності з медіанним заробітком випускників? Чи є суттєва різниця у медіанному заробітку між різними категоріями спеціальностей?

Відповідь: _____

**Висновки:** …

# Завдання для виконання лабораторної роботи

1. Завантажити набір даних «recent-grads.csv» у DataFrame Pandas.
2. Провести попередню обробку даних: видалити стовпці, які не потрібні для аналізу, та обробити пропущені значення, якщо вони є.
3. **Здобувач 1**. Розрахувати середній рівень працевлаштування (Employed / (Employed + Unemployed)) серед випускників різних спеціальностей для кожного варіанту.
4. **Здобувач 1**. Побудувати гістограму розподілу кількості студентів на спеціальності (Total) для кожного варіанту.
5. **Здобувач 1**. Знайти медіану кількості чоловіків (Men) на спеціальностях, де рівень працевлаштування вище середнього для даного варіанту.
6. **Здобувач 2**. Використовуючи дані, передані **Здобувачем 1**, побудувати діаграму розсіювання, що показує залежність між кількістю чоловіків (Men) та кількістю жінок (Women) на спеціальності для кожного варіанту. Додати на графік горизонтальну лінію, що відповідає медіані кількості чоловіків на спеціальностях, де рівень працевлаштування вище середнього, розраховану **Здобувачем 1**.
7. **Здобувач 2**. Використовуючи дані, передані Учасником 1, побудувати ящик з вусами для кількості повного робочого дня (Full_time) випускників різних категорій спеціальностей для кожного варіанту.
8. **Здобувач 2**. Розрахувати **secret key** як суму індексів (номер рядка) шести спеціальностей з найнижчим рівнем працевлаштування, використовуючи дані, передані **Здобувачем 1**.
9. Спільне завдання для **Здобувача 1** та **Здобувача 2**. Побудувати стовпчикову діаграму з накопиченням (stacked bar chart), яка показує розподіл кількості працевлаштованих на повний робочий день (Full_time), неповний робочий день (Part_time) та безробітних (Unemployed) для кожної спеціальності у вашому варіанті.

**Варіант 1.**
Категорія спеціальності: «Engineering»

**Варіант 2.**
Категорія спеціальності: «Business»

**Варіант 3.**
Категорія спеціальності: «Physical Sciences»

**Варіант 4.**
Категорія спеціальності: «Health»

**Варіант 5.**
Категорія спеціальності: «Computers & Mathematics»

**Варіант 6.**
Категорія спеціальності: «Agriculture & Natural Resources»

**Варіант 7.**
Категорія спеціальності: «Computers & Mathematics», «Physical Sciences»

**Варіант 8.**
Категорія спеціальності: «Arts»

**Варіант 9.**
Категорія спеціальності: «Social Science»

**Варіант 10.**
Категорія спеціальності: «Psychology & Social Work»

**Варіант 11.**
Категорія спеціальності: «Communications & Journalism»

**Варіант 12.**
Категорія спеціальності: «Law & Public Policy»

**Варіант 13.**
Категорія спеціальності: «Education»

**Варіант 14.**
Категорія спеціальності: «Humanities & Liberal Arts»

**Варіант 15.**
Категорія спеціальності: «Biology & Life Science»

**Варіант 16.**
Категорія спеціальності: «Engineering», «Business»

**Варіант 17.**
Категорія спеціальності: «Physical Sciences», «Health»

**Варіант 18.**
Категорія спеціальності: «Computers & Mathematics», «Agriculture & Natural Resources»

**Варіант 19.**
Категорія спеціальності: «Social Science», «Agriculture & Natural Resources»

**Варіант 20.**
Категорія спеціальності: «Social Science», «Psychology & Social Work»

**Варіант 21.**
Категорія спеціальності: «Communications & Journalism», «Law & Public Policy»

**Варіант 22.**
Категорія спеціальності: «Education», «Humanities & Liberal Arts»

**Варіант 23.**
Категорія спеціальності: «Biology & Life Science», «Engineering»

**Варіант 24.**
Категорія спеціальності: «Business», «Physical Sciences»

**Варіант 25.**
Категорія спеціальності: «Health», «Computers & Mathematics»

**Варіант 26.**
Категорія спеціальності: «Education», «Psychology & Social Work»

**Варіант 27.**
Категорія спеціальності: «Arts», «Social Science»

**Варіант 28.**
Категорія спеціальності: «Psychology & Social Work», «Communications & Journalism»

**Варіант 29.**
Категорія спеціальності: «Law & Public Policy», «Education»

**Варіант 30.**
Категорія спеціальності: «Humanities & Liberal Arts», «Biology & Life Science»

**Welcome to Google Colab**

https://colab.research.google.com/drive/1tb5_4PqsEefLDr6r2j29GdWSPnV86nDg?usp=sharing