НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО”
ФАКУЛЬТЕТ ІНФОРМАТИКИ ТА ОБЧИСЛЮВАЛЬНОЇ ТЕХНІКИ

Кафедра інформатики та програмної інженерії

Звіт до лабораторної роботи №1

з курсу

«Машине навчання»

студента 2 курсу
групи ІТ-02
Макарова Іллі Сергійовича

Викладач:
Оніщенко В.

Київ – 2022

Тема: Введення в data science

Завдання:

1. На сайті http://www.ukrstat.gov.ua/ оберіть дані які для Вас є цікавими, можна використати будь-який ресурс з відкритими даними, та завантажте дані

2. Знайдіть математическое сподівання, медіану, моду, дисперсію, середньоквадратичне відхилення (поясніть їх зміст)
3. Візуалізуйте завантажені дані за допомогою гістограми
4. Зробіть звіт про роботу, який включає:
   1.титульна сторінка з інформацією про виконавця, темою та номером лабораторної роботи,
   2. Постановку завдання
   3. Скрін коду та скрін результату виконання

Виконання:

Тож, датасет я взяв з Kaggle, *https://www.kaggle.com/themlphdstudent/gdp-by-country-2017*.

Це дані ВВП країн світу на 2017 рік.

Я не знаю можно ли использовать pandas в данной работе, но я с вашего позволения, файлик открою в нем, и достану данные, а дальше как вы показывали заюзаю **statistics**

In [3]:
```
df = pd.read_csv('data/GDP by Country in 2017.csv')
```

In [4]:
```
df.head()
```

	Rank	Country	GDP(in US$)	GDP (abbrev.)	GDP growth %	Population (2017)	GDP per capita(in US$)	Share of World GDP %
0	1	United States	19485394000000	$19.485 trillion	2.27	325084756	59939	24.08
1	2	China	12237700479375	$12.238 trillion	6.90	1421021791	8612	15.12
2	3	Japan	4872415104315	$4.872 trillion	1.71	127502725	38214	6.02
3	4	Germany	3693204332230	$3.693 trillion	2.22	82658409	44680	4.56
4	5	India	2650725335364	$2.651 trillion	6.68	1338676785	1980	3.28

In [15]:
```
df['GDP(in US$)'].isnull().sum()
```

```
0
```

In [21]:
```
gdp_raw = list(df['GDP(in US$)'])
gdp_raw
```

На скріншотах є мої коментарі, де я пояснюю свої дії. Я прочитав дані за допомогою pandas, та надалі я буду намагатись використовувати лише вбудованні інструменти пайтону.

In [14]:
```
print(f'Median of GDP in 2017 was:', statistics.median(gdp_raw))

try:
    print(f'Mode of GDP in 2017 was:', statistics.mode(gdp_raw))
except statistics.StatisticsError:
    print('Cannot calculate mode for GDP data')

print(f'Mean of GDP in 2017 was:', round(statistics.mean(gdp_raw), 2))
```

```
Median of GDP in 2017 was: 37353276059
Mode of GDP in 2017 was: 19485394000000
Mean of GDP in 2017 was: 421023956456.77
```

In [17]:
```
print(f'Dispersion of GDP in 2017 was:', statistics.pvariance(gdp_raw))
print(f'Standard diviation of GDP in 2017 was:', statistics.pstdev(gdp_raw))
```

```
Dispersion of GDP in 2017 was: 3.079113663813578e+24
Standard diviation of GDP in 2017 was: 1754740340852.0527
```

> Дисперсия выглядит какой-то очень большой, но если мы посмотрим в каких порядках у нас измеряется ВВП то все становится на свои места.

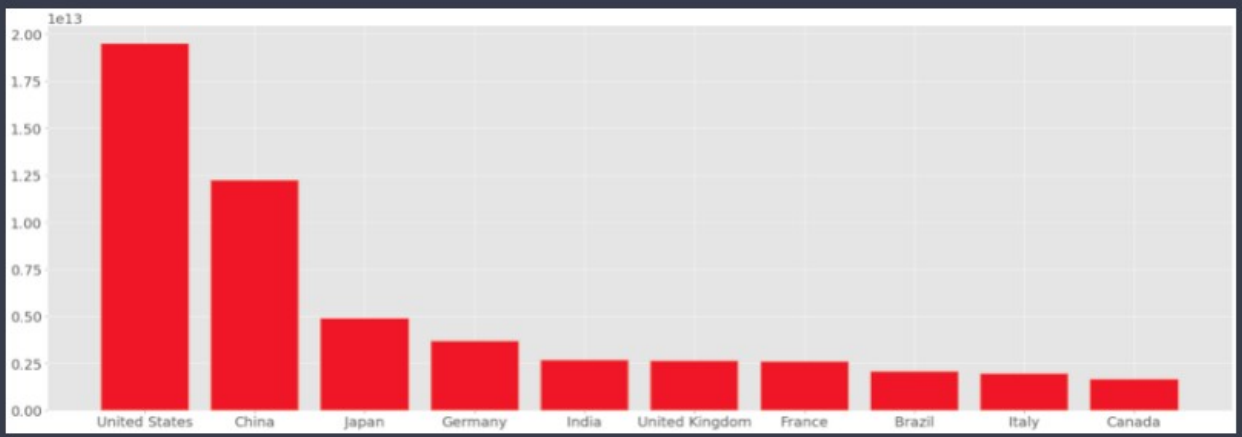
Тепер візуалізація:

> Я взял только первые 10 экономик мира, так как все красиво визуализировать довольно сложно

In [56]:
```
top_10_gdp = gdp_raw[:10]
top_10_countries = list(df['Country'])[:10]

plt.bar(top_10_countries, top_10_gdp)
plt.rcParams['figure.figsize'] = [30, 10]
plt.rcParams.update({'font.size': 25})

plt.show()
```



В задании было сделать гистограмму, поэтому я ща и ее сделаю, просто for legal reasons так сказать

In [61]:

```
plt.hist(gdp_raw, bins=50)
plt.xlabel('GDP')
plt.ylabel('Number of countries')

plt.show()
```