江苏大学京江学院

JIANGSU UNIVERSITY JINGJIANG COLLEGE

Python 编程与科学计算

题目：	对全球和中国互联网用户的数据分析与可视化
姓名：	马云骥
班级：	J 软件(嵌入)(专转本)2102
学号：	4211153047
教师：	王洪金

2024 年 5 月

对全球和中国互联网用户的数据分析与可视化

第一章 项目概述

1.1 项目背景

互联网是当今时代最重要和最有影响力的技术之一，它已经深刻地改变了人们的生活、工作、学习等方面。从购物、社交到获取信息和娱乐，互联网的普及和应用无处不在。互联网用户数据作为反映互联网发展水平和潜力的重要指标，能够帮助我们了解不同国家和地区在互联网领域的优势和劣势，以及存在的差异和不平衡。这些数据不仅能够提供当前的互联网普及情况，还可以帮助预测未来的发展趋势。

在过去的几十年里，全球互联网用户数量迅速增长，特别是在发展中国家，互联网的普及速度令人瞩目。中国作为全球最大的互联网市场，其互联网用户数量和使用情况引起了广泛关注。中国的互联网发展不仅仅体现在用户数量的增加，还体现在移动互联网的普及和宽带网络的覆盖范围上。这些变化不仅推动了经济的发展，也对社会文化产生了深远的影响。

然而，全球和中国的互联网用户数据也存在着一些问题和挑战。例如，不同地区之间的互联网普及率存在显著差异，特别是在发展中国家和发达国家之间。此外，互联网用户数据的增长并不总是均衡的，有些地区可能因为基础设施、经济水平或政策等原因，互联网普及率较低。通过对这些数据进行分析和可视化，我们可以更好地把握互联网领域的变化趋势和分布情况，识别出需要关注和改进的领域。

1.2 数据来源

数据来源于 Kaggle 上的 Global Internet Users 数据集[1]，该数据集包含了 1980-2020 年间关于全球互联网用户的信息，包括国家或地区名称、国家代码、年份、每 100 人的移动端互联网订阅数、互联网用户占总人口的比例、互联网用户数量以及每 100 人的宽带订阅数等信息。

1.3 程序功能

1.3.1 全球用户数据分析和可视化

- 绘制 2020 年各个国家地区的用户占比饼图和柱状图，展示全球互联网用户占比的分布情况和差异。

- 绘制 2020 年各国家地区互联网用户占比分布直方图，展示全球互联网用户占比的分布特征和偏态。
- 绘制 2020 年个国家地区互联网用户占比和移动互联网订阅量的散点图，并利用线性回归模型分析两者之间的相关性。
- 绘制每一年互联网用户的比例最大的三个国家地区名的词云，展示全球互联网领域的优势和影响力。

1.3.2 中国用户数据分析和可视化

- 对中国互联网用户数据进行分析和可视化，展示中国在互联网领域的发展水平和潜力。
- 利用多元线性回归模型预测中国互联网到 2050 年的总用户数。

1.4 所使用第三方库介绍

- NumPy 库[2]：用于进行数值计算，如数组、矩阵、向量等的创建、操作和运算。
- Pandas 库[3]：用于处理数据，如数据的读取、清洗、分组、聚合、合并等。
- Matplotlib 库[4]：用于绘制图形，如折线图、柱状图、饼图等，以及设置图形的样式、标题、标签等。
- Seaborn 库[5]：用于绘制图形，如直方图、散点图等，以及设置图形的主题、颜色等。
- WordCloud 库[6]：用于配置和生成词云，如设置词云的形状、大小、字体等。
- Sklearn 库[7]：用于进行线性回归分析，如创建线性回归模型、拟合数据、预测数据、评估模型等。

第二章 功能实现

2.1 数据读取和工具函数的实现

使用 read_csv 函数读取 Final.csv 文件，获取全球互联网用户信息，定义两个工具函数 set_seaborn_properties、get_2020_entities_dataframe。

```
# 从数据集中读取出所有的数据
global_users = pd.read_csv('Final.csv', delimiter=',', usecols=range(1, 8))

# 此函数用于配置每个 seaborn 图的主题，其中提供了一些默认配置，需要修改的配置在参数中提供，根据图的美观性和实用性需要改变参数
def set_seaborn_properties(context='talk', font_scale=0.8)

# 此函数用于获取所有 Entity 的 2020 年数据组成的 DataFrame
def get_2020_entities_dataframe()
```

2.2 全球用户每年的各项数据的分析与可视化

定义一个名为 global_internet_users_analysis 的函数，用于分析和可视化全球用户每年的各项数据。在这个函数中，使用 matplotlib 和 seaborn 库绘制了全球每年的互联网用户总数、全球每年每 100 人移动端互联网订阅数、互联网使用人数比例、每 100 人宽带订阅数的平均值等图形。

```
plt.subplot(2, 2, 1)  # 绘制子图
sns.lineplot(data=internet_users_sum_data, x='Year', y='sum')  # 折线图
plt.bar(column_mean.index, column_mean.values, color='cornflowerblue', width=0.6)  # 配合折线图使构图饱满
for column in ['Cellular Subscription', 'Internet Users(%)', 'Broadband Subscription']:
    plt.subplot(2, 2, i)
    i += 1
    sns.lineplot(data=max_data, x='Year', y='max', label=column + ' max', lw=2, linestyle=(0, (5, 1)))  # 最大值
    sns.lineplot(data=mean_data, x='Year', y='mean', label=column + ' mean', lw=3, linestyle=(0, (1, 1)))  # 平均值
```

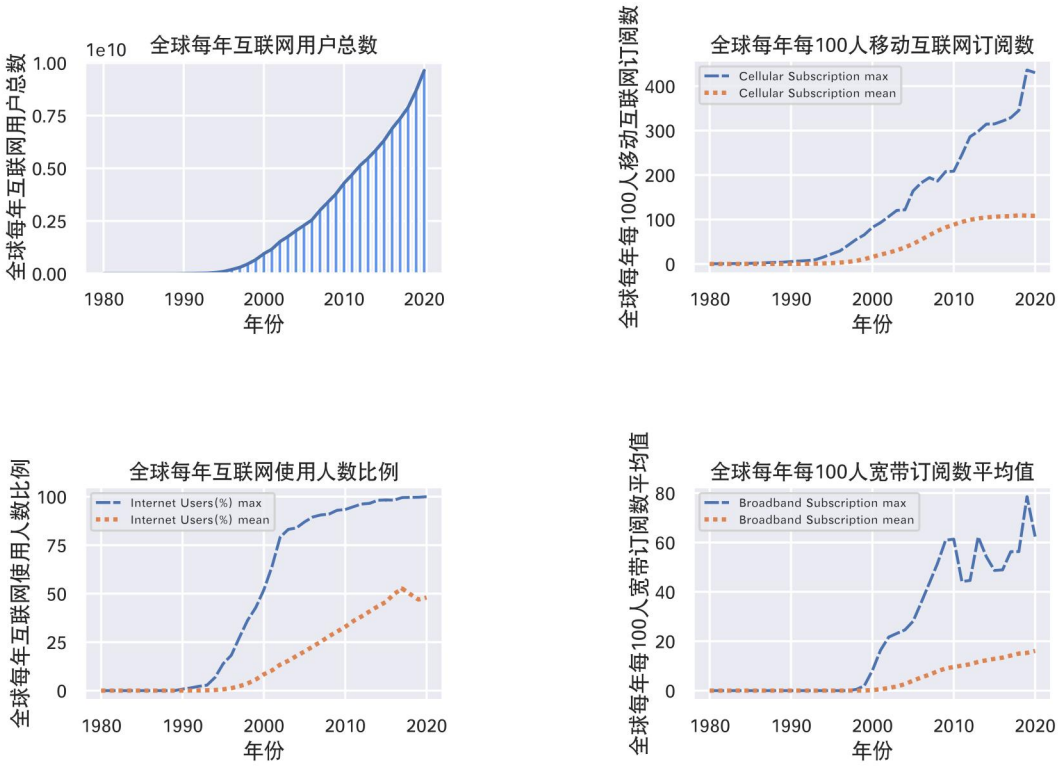

图 1　全球用户每年的各项数据的分析与可视化

从图 1 中可以看出：

- 全球互联网用户总数呈现出一个快速增长的趋势，尤其是在 2000 年之后，增长速度更加明显。这说明互联网技术的发展和普及，以及人们对互联网的需求和依赖都在不断增加。
- 全球每 100 人移动端互联网订阅数也呈现出一个快速增长的趋势，尤其是在 2005 年之后，增长速度更加明显。这说明移动设备的普及和便捷，以及人们对移动互联网的需求和偏好都在不断增加。
- 全球互联网使用人数比例也呈现出一个快速增长的趋势。这说明互联网已经成为人们生活、工作、学习等方面不可或缺的一部分，以及互联网的覆盖范围和接入方式都在不断扩大和改善。
- 全球每 100 人宽带订阅数呈现出一个缓慢增长的趋势，但在 2010 年之后，增长速度有所放缓。这说明宽带网络的发展和普及还有一定的空间和潜力，以及宽带网络的竞争力和吸引力可能受到了移动网络的影响。

2.3 2020 年各个国家地区的用户占比饼图和柱状图绘制

定义了一个名为 entities_2020_internet_users_percentage_pie_bar 的函数，用于绘制 2020 年各个国家地区的用户占比饼图和柱状图。在这个函数中，首先获取 2020 年国家地区用户数量最多的 10 组数据，其他数据用 other 代替的数据封装成的 DataFrame，使用 matplotlib 和 seaborn 库绘制饼图和柱状图。

```
# 只筛选用户数量最多的10组数据，其他数据用`other`代替
entity_2020_df.sort_values(by='No. of Internet Users', axis=0, ascending=False, inplace=True)
processed_data = pd.concat([entity_2020_df.head(10), other_df], axis=0, join='outer')

# 绘制饼图
plt.pie(processed_data['No. of Internet Users'], labels=processed_data.index, explode=explode_arr, labeldistance=1.1,
        autopct='%2.1f%%', pctdistance=0.9, shadow=True)

# 绘制柱状图
sns.barplot(data=data, x='Entity', y='Percent')
```

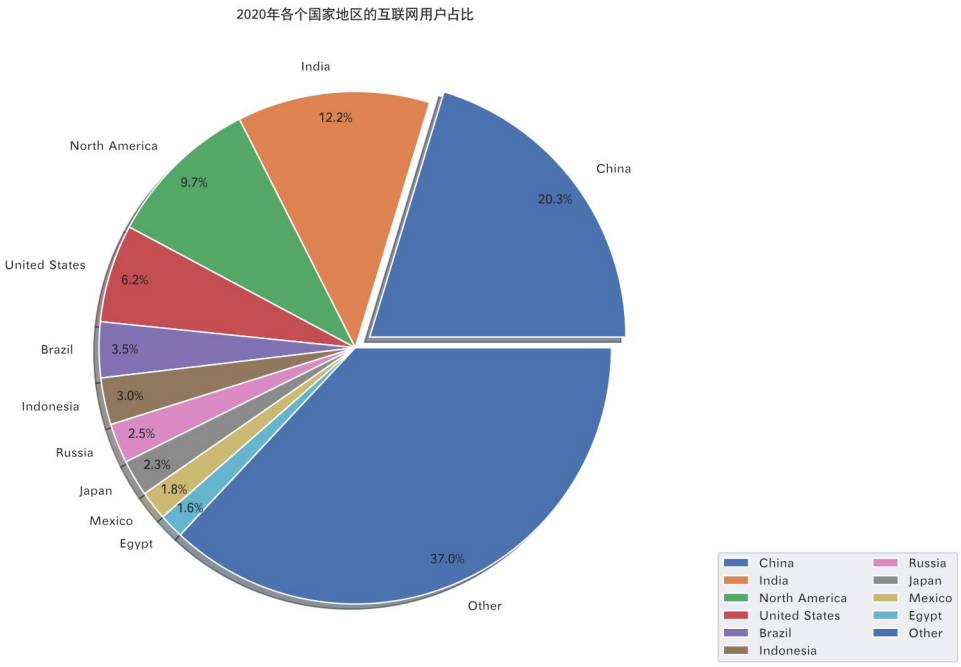

图 2　2020 年各个国家地区的互联网用户占比饼图

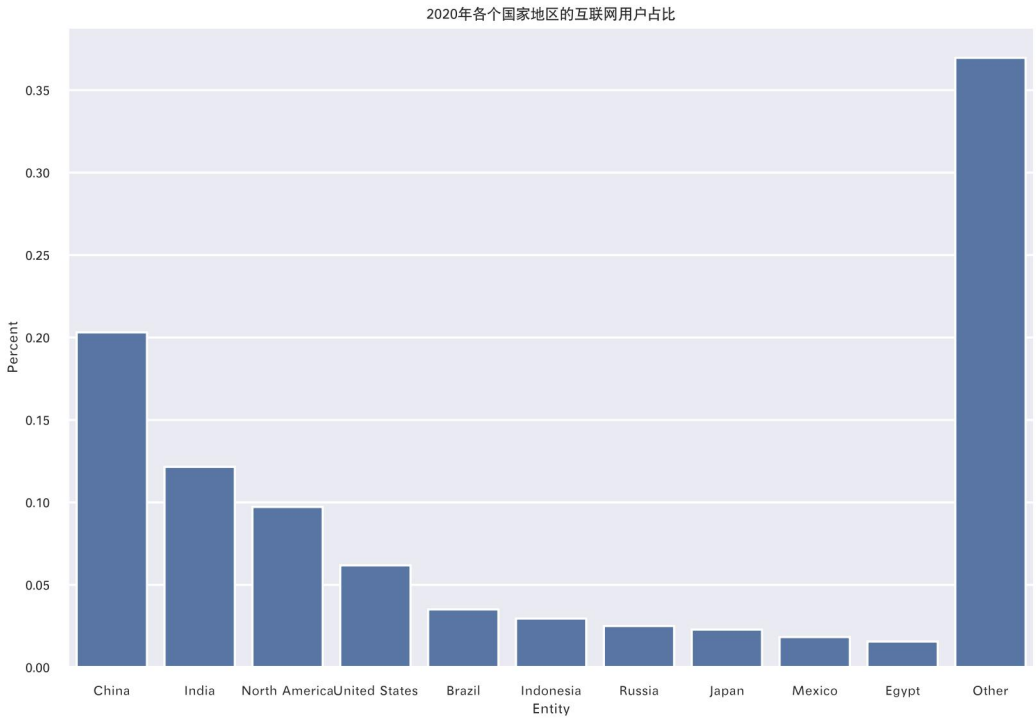

图 3　2020 年各个国家地区的互联网用户占比柱状图

从图 2、图 3 中可以看出：

- 在 2020 年，全球互联网用户占比最高的国家地区是中国，占比达到了 20.3%，远高于其他国家地区。这说明中国在互联网领域有着巨大的市场规模和潜力，以及中国在互联网技术、应用、服务等方面有着较强的竞争力和影响力。
- 在 2020 年，全球互联网用户占比第二高的国家地区是印度，占比为 12.1%，但与中国相比还有较大的差距。这说明印度在互联网领域也有着较大的市场规模和潜力，但与中国相比还有较大的发展空间和挑战。
- 在 2020 年，全球互联网用户占比第三高的国家地区是南美，占比为 9.7%，与印度相比较为接近。这说明美国在互联网领域也有着较大的市场规模和潜力，但与中国相比也有较大的差距。
- 在 2020 年，全球互联网用户占比排名前十的国家地区还包括巴西、印度尼西亚、俄罗斯、日本、墨西哥和埃及。这些国家地区的占比都在 1.5%到 3.6%之间，相对较低。这说明这些国家地区在互联网领域还有较大的发展空间和潜力，但也面临着较大的挑战和竞争。
- 在 2020 年，全球互联网用户占比排名第十一以后的国家地区的占比加起来只有约 37%，都远低于中国的占比。这说明这些国家地区在互联网领域还有较大的不平衡和差距，需要加强互联网技术的普及和提升。

2.4 2020 年各国家地区互联网用户占比分布直方图

定义了一个名为 entities_2020_internet_users_percentage_distribution_histogram 的函数，用于绘制 2020 年各国家地区互联网用户占比分布直方图。在这个函数中，首先获取 2020 年国家地区数据封装成的 DataFrame，然后使用 seaborn 库绘制了直方图。

```
# 获取数据
data = pd.DataFrame({'Entity': internet_users_percentage_sr.index,
'Percent':internet_users_percentage_sr.values})


# 绘制直方图
sns.histplot(data, x='Percent')
```

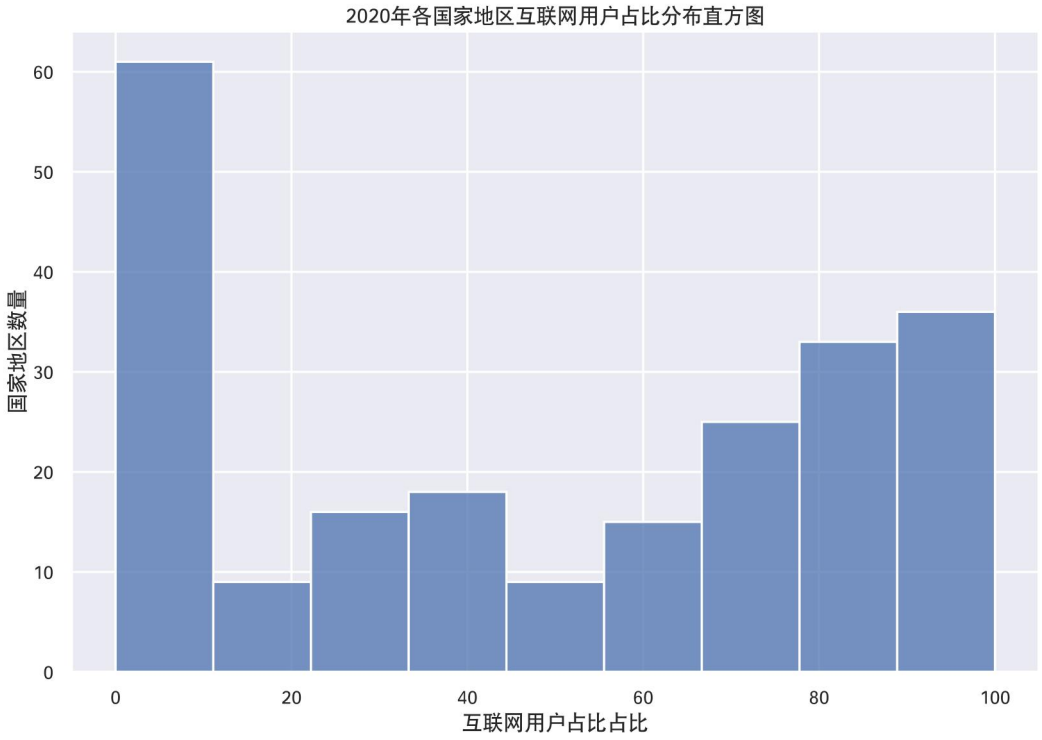

图 4　2020 年各个国家地区的互联网用户占比分布直方图

从图 4 中可以看出：

- 在 2020 年，各国家地区互联网用户占比的分布呈现出一个右偏态的分布，大部分国家地区的互联网用户占比都集中在较低的范围内，而少数国家地区的互联网用户占比则达到了较高的水平。
- 在 2020 年，各国家地区互联网用户占比的最高值为 100%，最低值为 0%，平均值为 47.9%，中位数为 53.9%，标准差为 36.1%。这说明各国家地区互联网用户占比存在着较大的差异和不平衡，以及部分国家地区的互联网发展水平还有较大的提升空间。

2.5 2020 年个国家地区互联网用户占比和移动互联网订阅量的散点图

定义了一个名为 entities_2020_internet_users_percentage_distribution_scatter 的函数，用于绘制 2020 年个国家地区互联网用户占比和移动互联网订阅量的散点图。在这个函数中，首先获取 2020 年国家地区数据封装成的 DataFrame，然后使用 seaborn 库绘制了散点图，并利用 sklearn 库中的线性回归模型分析两者之间的关系。

```
# 绘制散点图
sns.scatterplot(data=entity_2020_df, x='Internet Users(%)', y='Cellular Subscription', palette='husl', hue='Entity', legend=None)  # 根据地区设置 hue 参数，使颜色丰富


# 一元线性回归分析两者关系
x = entity_2020_df[['Internet Users(%)']]
model_1 = linear_model.LinearRegression()
model_1.fit(x, entity_2020_df[['Cellular Subscription']])
data = pd.DataFrame({'x': x['Internet Users(%)'], 'pred_y': [x[0] for x in model_1.predict(x)]})
sns.lineplot(data=data, x='x', y='pred_y')
```

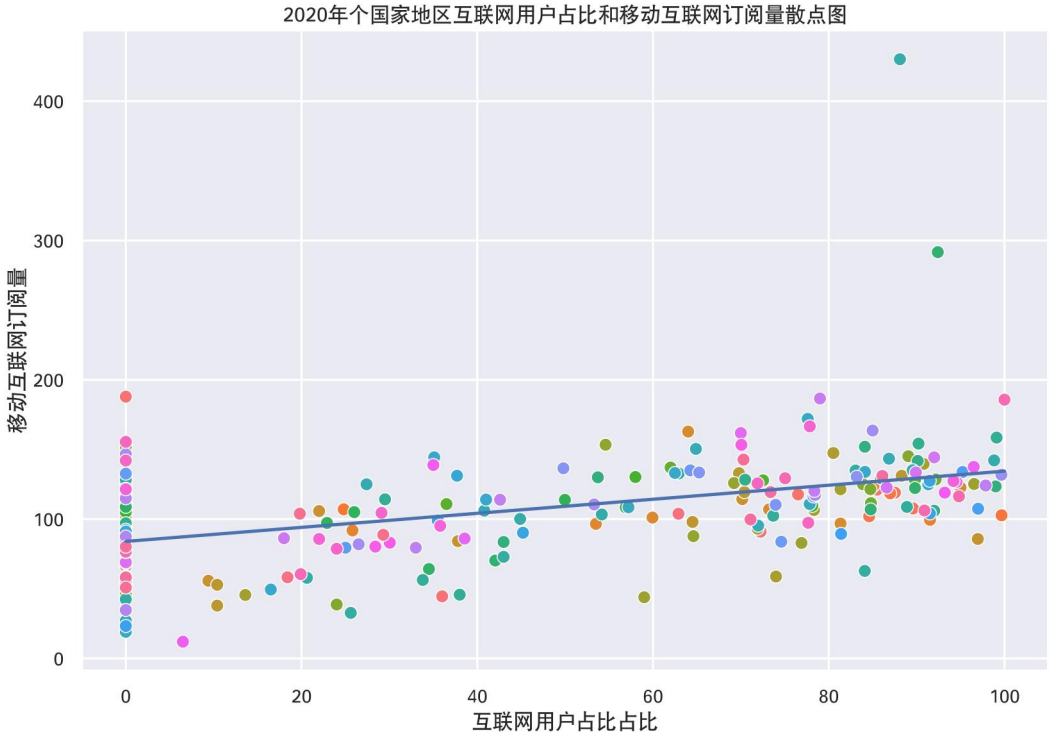

图 5　2020 年个国家地区互联网用户占比和移动互联网订阅量散点图及线性回归拟合

从图 5 中可以看出：

- 在 2020 年，各国家地区互联网用户占比和移动互联网订阅量呈现出一个正相关的关

系，即互联网用户占比越高的国家地区，移动互联网订阅量也越高，反之亦然。这说明互联网用户占比和移动互联网订阅量是两个相互影响和促进的指标，反映了一个国家地区的互联网发展水平和便捷程度。

2.6 用每一年互联网用户的比例最大的三个国家地区名生成词云

定义了一个名为 draw_internet_users_percentage_annual_top_3_wordcloud 的函数，用于绘制每一年互联网用户的比例最大的三个国家地区名生成词云。在这个函数中，首先获取每一年互联网用户的比例最大的三个国家地区名，然后使用 wordcloud 库绘制词云。

```
text = ''
year_groups = global_users.groupby('Year')
# 获取每一年互联网用户的比例最大的三个国家地区名数据
for year, year_df in year_groups:
    year_df.sort_values(by='Internet Users(%)', ascending=False, inplace=True)
    top_3 = year_df.head(3)
    entities = top_3['Entity']
    # 数据处理
    for entity in entities:
        if len(entity.split()) > 1:
            text += entity.replace(' ', '_') + ' '
            # 将名字中含有空格的国家地区名中的空格替换成下划线_，避免一个名字被拆分成多个单词
        else:
            text += entity + ' '

# 绘制词云
wc = WordCloud(max_words=100, width=800, height=400, background_color='White',
               max_font_size=150, stopwords=STOPWORDS, margin=5, scale=1.5)
wc.generate(text)
plt.imshow(wc)
plt.axis("off")
plt.show()
```

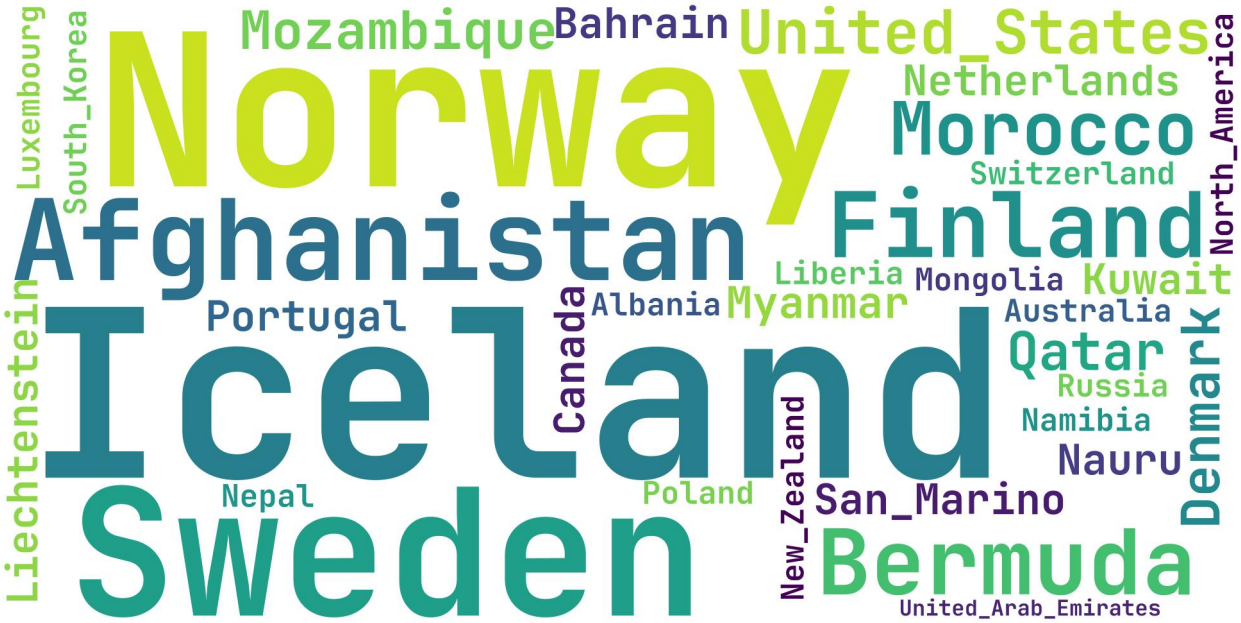

图 6　每年互联网用户的比例最大的国家地区名词云

从图 6 中可以看出：

- 在 1980-2020 年间，出现频率最高的国家地区名是 Iceland、Norway 和 Sweden。这说明这些国家地区在互联网领域有着长期的较高的发展水平和优势，以及较高的人口普及率和接入率。
- 在 1980-2020 年间，出现频率较高的国家地区名还有 Bermuda、Denmark、Finland、Moracco、Afghanistan 和 United_States 等。这说明这些国家地区在互联网领域也有着长期的较高的发展水平和优势，以及较高的人口普及率和接入率。
- 在 1980-2020 年间，出现频率较低或没有出现的国家地区名有 China、India、Brazil、Indonesia 等。结合 2.4 中的结果，这些国家地区有些是后起之秀，如中国和印度，有些则在互联网领域还有较大的发展空间和潜力，如巴西和印度尼西亚。

2.7 对中国互联网用户数据的分析与可视化

最后，我们定义了一个名为 chinese_users_analysis 的函数，用于对中国互联网用户数据进行分析和可视化。首先通过切片获取中国互联网用户信息。然后使用 matplotlib 和 seaborn 库绘制了各项指标的数值图和增长率图，并利用 sklearn 库中的多元线性回归模型预测中国互联网到 2050 年的总用户数。

```
# 基本信息的折线图
sns.lineplot(data=chinese_users, x='Year', y='No. of Internet Users', label='数量（单位：千万人）', lw=3)
sns.lineplot(data=chinese_users, x='Year', y='Internet Users(%)', label='占人口的比例', lw=3)
```

```
sns.lineplot(data=chinese_users, x='Year', y='Cellular Subscription', label='移动互联网订阅每一百人比例', lw=3)
sns.lineplot(data=chinese_users, x='Year', y='Broadband Subscription', label='宽带每一百人订阅比例', lw=3)
```

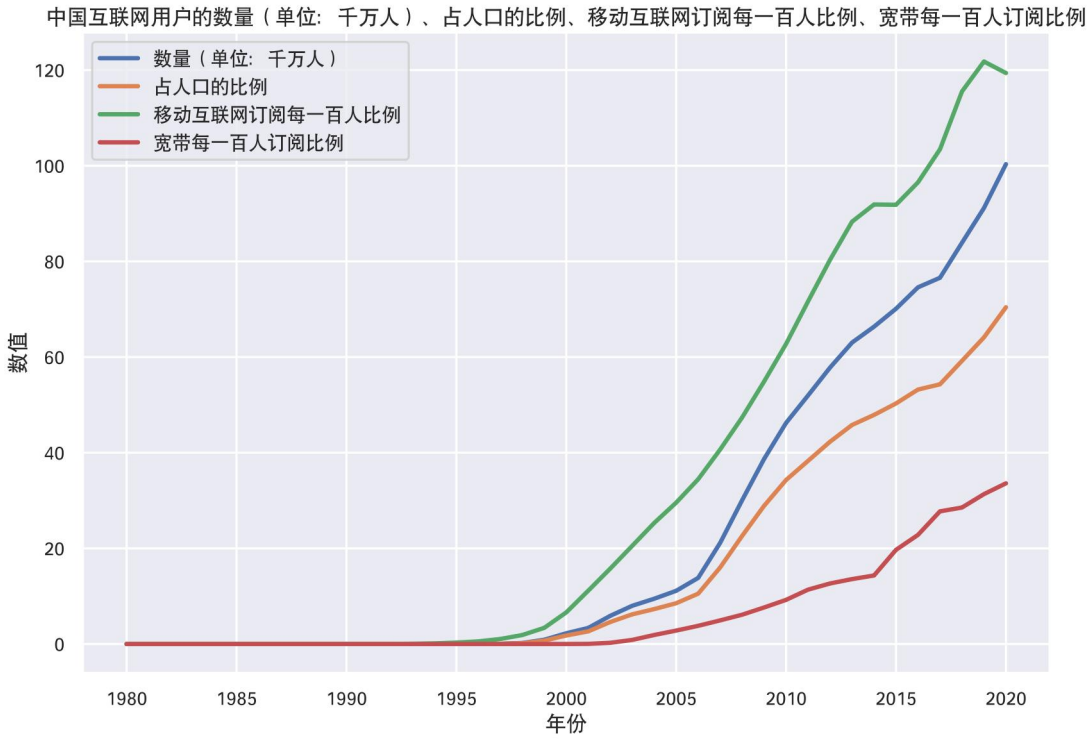

图 7　基本信息的折线图

从图 7 中可以看出：

- 中国互联网用户数量呈现出一个快速增长的趋势，尤其是在 2000 年之后，增长速度更加明显。这说明中国在互联网领域有着巨大的市场规模和潜力，以及中国在互联网技术、应用、服务等方面有着较强的竞争力和影响力。
- 中国互联网用户占人口比例也呈现出一个快速增长的趋势，尤其是在 2005 年之后，增长速度更加明显。这说明中国在互联网领域有着较高的普及率和接入率，以及中国在互联网领域的需求和依赖都在不断增加。
- 中国移动互联网订阅每一百人比例也呈现出一个快速增长的趋势，尤其是在 2005 年之后，增长速度更加明显。这说明中国在移动设备领域有着较高的普及率和便捷程度，以及中国在移动互联网领域的需求和偏好都在不断增加。
- 中国宽带每一百人订阅比例呈现出一个缓慢增长的趋势，但在 2017 年之后，增长速度有所放缓。这说明中国在宽带网络领域还有一定的空间和潜力，以及中国在宽带网络领域的竞争力和吸引力可能受到了移动网络的影响。

```
# 计算各项增长率
rows = len(chinese_users.index)
for i in range(rows - 1):
    chinese_users.loc[:, 'increase of No. of Internet Users'].iloc[i + 1] = 0 if
chinese_users.iloc[i]['No. of Internet Users'] == 0 else (chinese_users.iloc[i +
1].loc['No. of Internet Users'] - chinese_users.iloc[i]['No. of Internet Users']) /
chinese_users.iloc[i]['No. of Internet Users']
......


# 绘制图形
sns.lineplot(data=chinese_users, x='Year', y='increase of No. of Internet Users', lw=4,
label='数量（单位：千万人）增长率')
......
```

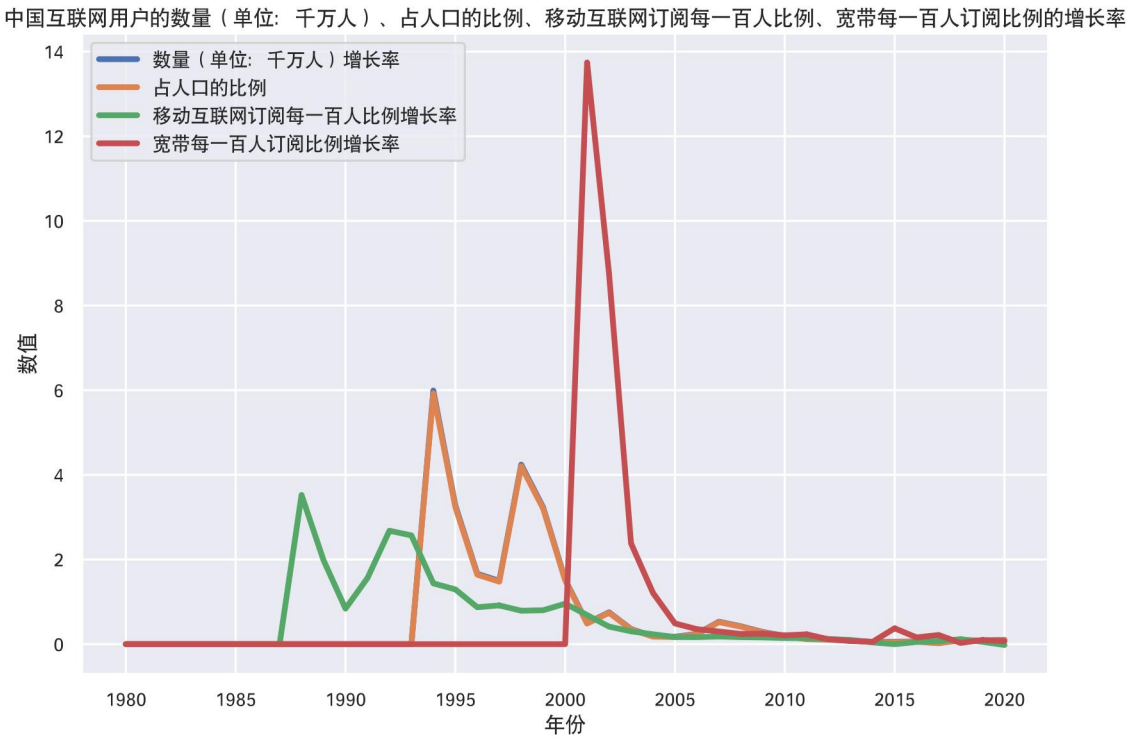

图 8　各项数据的增长率

从图 8 中可以看出：

- 中国互联网用户数量增长率和占人口比例增长率基本为同比增长，在 1993-2005 年之间增长较快之后趋于平稳，这说明中国在互联网领域已经达到了一个较高的发展水平，中国在互联网正在稳步发展。
- 中国移动互联网订阅每一百人比例增长率呈现在 1987-2005 年之间增长较快之后趋于平稳，尤其是在 2010 年之后，这说明中国在移动设备领域已经达到了一个较高的普及率和便捷程度，以及中国在移动互联网领域已形成良好发展趋势。

- 中国宽带每一百人订阅比例增长率在 2000-2002 年飞速增长之后又下降并趋于平稳。这说明中国在宽带网络领域有一定的空间和潜力，以及中国在宽带网络领域有着较强的竞争力和吸引力。

```
# 散点图
sns.scatterplot(data=chinese_users, x='Year', y='No. of Internet Users')

# 三元线性回归拟合
poly_reg = PolynomialFeatures(degree=3)
......
model_2.fit(x_m, chinese_users[['No. of Internet Users']])
data = pd.DataFrame({'x': x['Year'], 'pred_y': [x[0] for x in model_2.predict(x_m)]})

# 绘制折线图
sns.lineplot(data=data, x='x', y='pred_y')
```

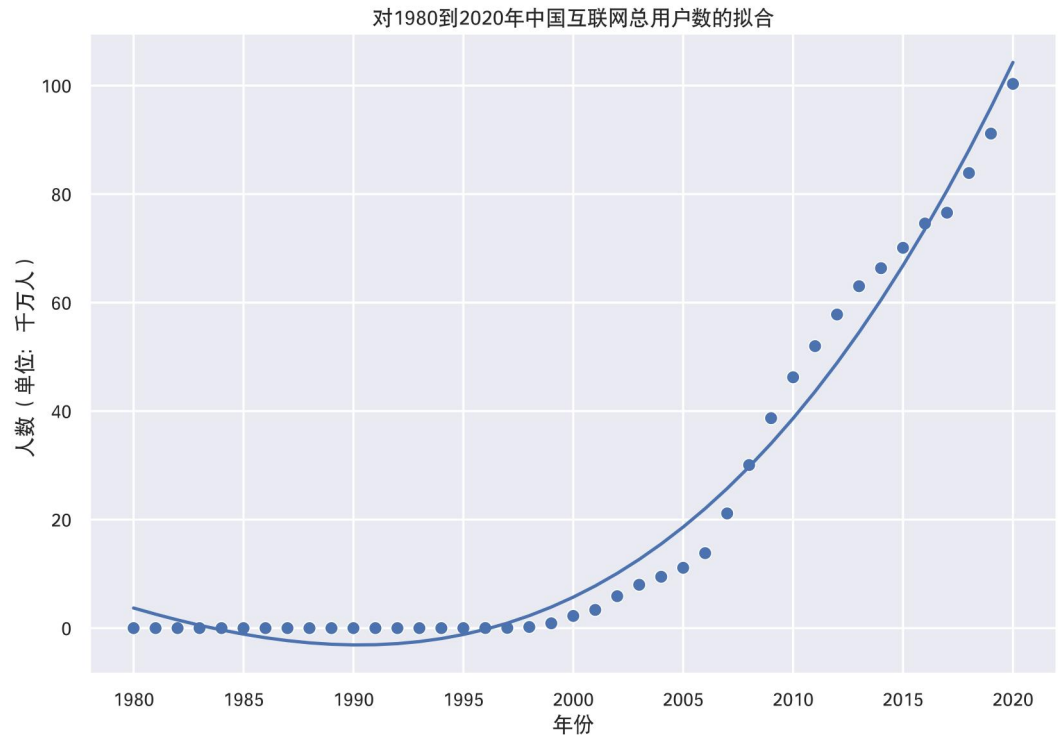

图 9　对 1980 到 2020 年中国互联网总用户数的拟合

从图 9 中可以看出：

- 我们使用 sklearn 库中的多元线性回归模型对 1980 到 2020 年中国互联网总用户数进行了拟合，得到了一个拟合曲线。这个曲线可以用来描述中国互联网总用户数随时间的变化规律，以及评估拟合效果和意义。

```
# 预测
pred_x = pd.DataFrame(np.arange(1980, 2031), columns=['Year'])
pred_x_m = poly_reg.fit_transform(pred_x)

# 绘图
plt.plot(pred_x, model_2.predict(pred_x_m))
```

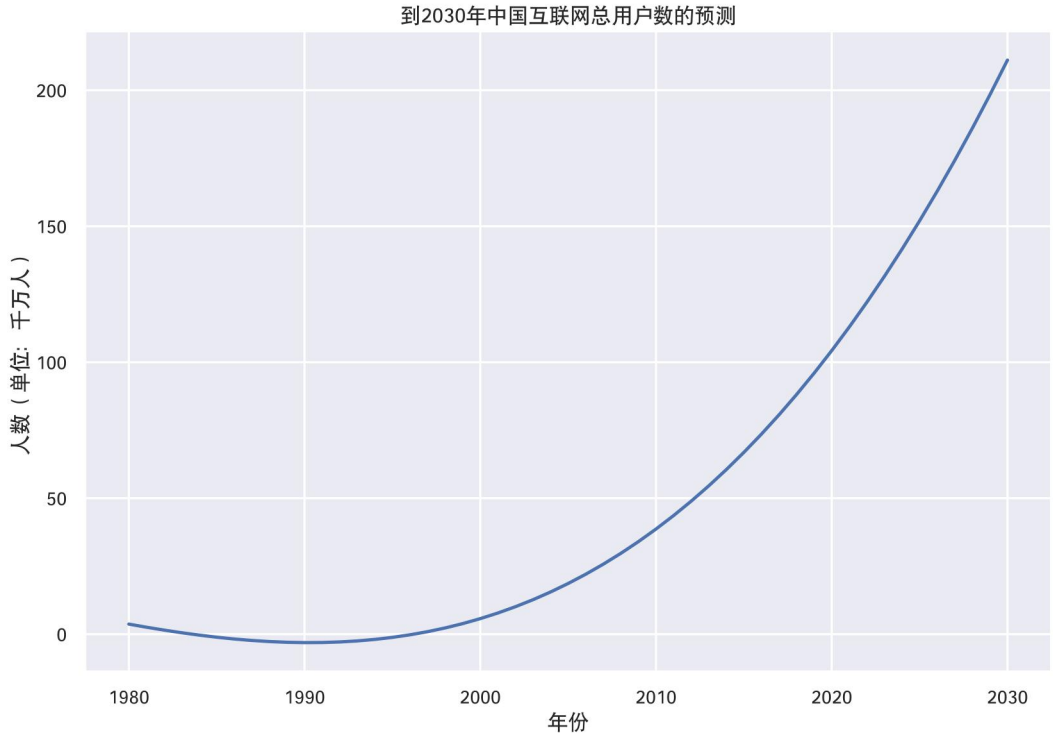

图 10　到 2030 年中国互联网总用户数的预测

从图 10 中可以看出：

- 到 2030 年中国互联网总用户数将达到 21.1 亿，此预测具有一定的合理性，因为中国互联网用户的各项指标都在飞速增长，中国互联网具有很大的发展潜力与发展活力。
- 但是此预测也具有一定的局限性，这个预测仅仅只用了一个数据集，没有考虑中国具体国情，仅仅依托数量的线性增长来分析是不合理的，需要更加高级的模型，并且要兼顾中国人口老龄化问题，结合中国人口增长速度来进一步分析，这样的预测效果会更好。

第三章 总结体会

在本项目中，我通过分析全球和中国互联网用户数据，取得了编程思想和编程技能方面的显著进步。

首先，我深刻体会到数据驱动思维的重要性。数据不仅仅是静态的数字，更是洞察和决策的基础。我学会了系统化地处理数据，从清洗、预处理到分析和可视化，每一步都注重规范化操作，以确保数据的准确性和有效性。同时，通过面对具体问题进行编程，我学会了将抽象的问题具体化，并找到针对性的解决方案。这种面向问题的编程思想帮助我提高了分析和解决实际问题的能力。

在实际操作中，我对 Python 语言及其第三方库有了更深入的了解。通过大量使用 numpy、pandas、matplotlib、seaborn 等库，我提升了数据处理和可视化的能力。同时，模块化编程和函数复用的实践使我的代码更加简洁、高效和可维护。通过定义和调用函数，我减少了冗余代码，提高了程序的可读性和复用性。

总的来说，通过本项目，我不仅提升了数据分析和编程技能，还建立了系统化、面向问题的编程思想。未来，我将继续深化这些技能，结合更多的实际问题进行探索和实践，不断提升自己的技术水平和解决问题的能力。我相信，通过持续的学习和实践，我能够在数据分析和编程领域取得更大的进步和成就。

附件 1 参考文献

[1] Ashishraut64. Global Internet Users [DS/OL]. Kaggle, 2023[2024-05-20]. https://www.kaggle.com/datasets/ashishraut64/internet-users.

[2] NumPy. [EB/OL]. NumPy, [2024-05-20]. https://numpy.org/.

[3] Pandas. [EB/OL]. PyData, [2024-05-20]. https://pandas.pydata.org/.

[4] Matplotlib. [EB/OL]. Matplotlib, [2024-05-20]. https://matplotlib.org/.

[5] Seaborn. [EB/OL]. PyData, [2024-05-20]. https://seaborn.pydata.org/.

[6] Amueller. WordCloud [EB/OL]. GitHub, [2024-05-20]. https://amueller.github.io/word_cloud/.

[7] Scikit-learn. [EB/OL]. Scikit-learn, [2024-05-20]. https://scikit-learn.org/stable/.

附件 2 程序代码

```
import os
import sys
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
from wordcloud import STOPWORDS, WordCloud
from sklearn import linear_model
from sklearn.preprocessing import PolynomialFeatures


# 环境和参数的配置
def set_seaborn_properties(context='talk', font_scale=0.8):
    # 根据不同操作系统选用不同的字体
    if sys.platform == 'win32':  # Windows 下使用微软雅黑
        defaultFont = 'Microsoft YaHei'
    elif sys.platform == 'darwin':  # MacOS 下使用黑体
        defaultFont = 'Hei'
    else:  # Linux 下使用 Noto Sans CJK JP
        defaultFont = 'Noto Sans CJK JP'  # JP 为日本地区的默认中文字形
        # defaultFont = 'Noto Sans CJK SC'  SC 为中国大陆的默认中文字形，但是不知为何在本人电脑无法调用 (Ubuntu 24.04)
    sns.set_theme(context=context, font=defaultFont, font_scale=font_scale,
                  rc={'axes.unicode_minus': False,
                      'figure.figsize': (12, 8),
                      'figure.dpi': 150})


# 获取 2020 年国家地区数据封装成的 DataFrame
def get_2020_entities_dataframe():
    entity_group = global_users.groupby('Entity')
    entity_2020_df = pd.DataFrame()
    for entity, entity_df in entity_group:
        if entity == 'World':
            continue
        entity_2020 = entity_df[entity_df['Year'] == 2020]
        entity_2020_df = pd.concat([entity_2020_df, entity_2020], join='outer', axis=0)
    return entity_2020_df.set_index('Entity')


# 全球用户每年的各项数据的分析与可视化
```

```
def global_internet_users_analysis():
    set_seaborn_properties()
    # 全球每年的互联网用户总数分析与可视化：
    plt.subplots_adjust(hspace=1, wspace=0.7)
    year_groups = global_users.groupby('Year')
    internet_users_groups = year_groups['No. of Internet Users']
    column_mean = internet_users_groups.sum()
    internet_users_sum_data = pd.DataFrame({'Year': column_mean.index,
                                            'sum': column_mean.values})
    plt.subplot(2, 2, 1)
    plt.title('全球每年互联网用户总数')
    plt.xlabel('年份')
    plt.ylabel('全球每年互联网用户总数')
    sns.lineplot(data=internet_users_sum_data, x='Year', y='sum')
    plt.bar(column_mean.index,      column_mean.values,      color='cornflowerblue',
width=0.6)


    # 全球每年每100人移动端互联网订阅数、互联网使用人数比例、每100人宽带订阅数的平均值分析与可视
化：
    year_groups = global_users.groupby('Year')
    title_mapper = {'Cellular Subscription': '全球每年每100人移动互联网订阅数',
                    'Internet Users(%)': '全球每年互联网使用人数比例',
                    'Broadband Subscription': '全球每年每100人宽带订阅数平均值'}
    i = 2
    for  column  in  ['Cellular  Subscription',  'Internet  Users(%)',  'Broadband
Subscription']:
        plt.subplot(2, 2, i)
        i += 1
        internet_users_groups = year_groups[column]
        column_mean = internet_users_groups.mean()
        column_max = internet_users_groups.max()
        max_data = pd.DataFrame({'Year': column_max.index, 'max': column_max.values})
        mean_data     =     pd.DataFrame({'Year':     column_mean.index,     'mean':
column_mean.values})
        plt.title(title_mapper[column])
        plt.xlabel('年份')
        plt.ylabel(title_mapper[column])
        sns.lineplot(data=max_data, x='Year', y='max', label=column + ' max', lw=2,
linestyle=(0, (5, 1)))
        sns.lineplot(data=mean_data, x='Year', y='mean', label=column + ' mean', lw=3,
linestyle=(0, (1, 1)))
        plt.legend(loc='upper left', prop={'size': 8.5})
    plt.savefig(svgSavePath + '全球用户每年的各项数据的分析与可视化.svg')
    plt.savefig(pngSavePath + '全球用户每年的各项数据的分析与可视化.png')
    plt.show()
```

```
# 2020 年各个国家地区的用户占比饼图和柱状图绘制
def entities_2020_internet_users_percentage_pie_bar():
    entity_2020_df = get_2020_entities_dataframe()
    entity_2020_df['No. of Internet Users'] /= entity_2020_df['No. of Internet Users'].sum()

    # 只筛选用户数量最多的10组数据，其他数据用`other`代替
    entity_2020_df.sort_values(by='No. of Internet Users', axis=0, ascending=False, inplace=True)
    other = entity_2020_df.iloc[10:].loc[:, 'No. of Internet Users'].sum()
    other_df = pd.DataFrame(data={'': {'Entity': 'Other', 'No. of Internet Users': other}}).T
    other_df.set_index('Entity', inplace=True)
    processed_data = pd.concat([entity_2020_df.head(10), other_df], axis=0, join='outer')

    # 绘制饼图
    set_seaborn_properties(context='notebook', font_scale=0.8)
    explode_arr = np.zeros(shape=(11))
    explode_arr[0] = 0.07
    plt.axes(aspect=1)
    plt.title('2020 年各个国家地区的互联网用户占比')
    plt.pie(processed_data['No. of Internet Users'], labels=processed_data.index, explode=explode_arr,
            labeldistance=1.1, autopct='%2.1f%%', pctdistance=0.9, shadow=True)
    plt.legend(loc='lower right', bbox_to_anchor=(0.5, 0., 0.95, 0.5), ncols=2)
    plt.savefig(svgSavePath + '2020 年各个国家地区的互联网用户占比饼图.svg')
    plt.savefig(pngSavePath + '2020 年各个国家地区的互联网用户占比饼图.png')
    plt.show()

    # 绘制柱状图
    set_seaborn_properties(font_scale=0.56)
    plt.rcParams['figure.dpi'] = 300
    data = pd.DataFrame({'Entity': processed_data.index, 'Percent': processed_data['No. of Internet Users']})
    plt.title('2020 年各个国家地区的互联网用户占比')
    sns.barplot(data=data, x='Entity', y='Percent')
    plt.savefig(svgSavePath + '2020 年各个国家地区的互联网用户占比柱状图.svg')
    plt.savefig(pngSavePath + '2020 年各个国家地区的互联网用户占比柱状图.png')
    plt.show()


# 2020 年各国家地区互联网用户占比分布直方图
```

```
def entities_2020_internet_users_percentage_distribution_histogram():
    set_seaborn_properties(font_scale=0.8)
    entity_2020_df = get_2020_entities_dataframe()
    internet_users_percentage_sr = entity_2020_df['Internet Users(%)']
    plt.title('2020年各国家地区互联网用户占比分布直方图')
    plt.xlabel('互联网用户占比占比')
    plt.ylabel('国家地区数量')
    data = pd.DataFrame({'Entity': internet_users_percentage_sr.index, 'Percent': internet_users_percentage_sr.values})
    sns.histplot(data, x='Percent')
    plt.savefig(svgSavePath + '2020年各国家地区互联网用户占比分布直方图.svg')
    plt.savefig(pngSavePath + '2020年各国家地区互联网用户占比分布直方图.png')
    plt.show()


# 2020年个国家地区互联网用户占比和移动互联网订阅量的散点图
def entities_2020_internet_users_percentage_distribution_scatter():
    set_seaborn_properties()
    entity_2020_df = get_2020_entities_dataframe()
    plt.title('2020年个国家地区互联网用户占比和移动互联网订阅量散点图')
    plt.xlabel('互联网用户占比占比')
    plt.ylabel('移动互联网订阅量')
    sns.scatterplot(data=entity_2020_df,    x='Internet    Users(%)',    y='Cellular Subscription',
                    palette='husl', hue='Entity', legend=None)

    # 利用线性回归分析两者关系
    x = entity_2020_df[['Internet Users(%)']]
    model_1 = linear_model.LinearRegression()
    model_1.fit(x, entity_2020_df[['Cellular Subscription']])
    data = pd.DataFrame({'x': x['Internet Users(%)'], 'pred_y': [x[0] for x in model_1.predict(x)]})
    sns.lineplot(data=data, x='x', y='pred_y')
    plt.savefig(svgSavePath + '2020年个国家地区互联网用户占比和移动互联网订阅量散点图及线性回归拟合.svg')
    plt.savefig(pngSavePath + '2020年个国家地区互联网用户占比和移动互联网订阅量散点图及线性回归拟合.png')
    plt.show()


# 用每一年互联网用户的比例最大的三个国家地区名生成词云
def draw_internet_users_percentage_annual_top_3_wordcloud():
    text = ''
    year_groups = global_users.groupby('Year')
```

```
    # 获取每一年互联网用户的比例最大的三个国家地区名数据
    for year, year_df in year_groups:
        year_df.sort_values(by='Internet Users(%)', ascending=False, inplace=True)
        top_3 = year_df.head(3)
        entities = top_3['Entity']
        for entity in entities:
            if len(entity.split()) > 1:
                text += entity.replace(' ', '_') + ' '
                # 将名字中含有空格的国家地区名中的空格替换成下划线_，避免一个名字被拆分成多个单词
            else:
                text += entity + ' '
    wc = WordCloud(font_path='JetBrainsMono-ExtraBold.ttf', max_words=100, width=800,
height=400, background_color='White',
                   max_font_size=150, stopwords=STOPWORDS, margin=5, scale=1.5)
    wc.generate(text)
    plt.title('每年互联网用户的比例最大的国家地区名词云')
    plt.imshow(wc)
    plt.axis("off")
    with open(svgSavePath + '每年互联网用户的比例最大的国家地区名词云.svg', 'w') as f:
        f.write(wc.to_svg())
    wc.to_file(pngSavePath + '每年互联网用户的比例最大的国家地区名词云.png')
    plt.show()


# 对中国互联网用户数据的分析与可视化
def chinese_users_analysis():
    # 绘制各项指标的数值图
    set_seaborn_properties()
    pd.options.mode.chained_assignment = None
    plt.title('中国互联网用户的数量（单位：千万人）、占人口的比例、移动互联网订阅每一百人比例、宽
带每一百人订阅比例')
    plt.xlabel('年份')
    plt.ylabel('数值')
    chinese_users.loc[:, 'No. of Internet Users'] /= 10000000
    sns.lineplot(data=chinese_users, x='Year', y='No. of Internet Users', label='数
量（单位：千万人）', lw=3)
    sns.lineplot(data=chinese_users, x='Year', y='Internet Users(%)', label='占人口
的比例', lw=3)
    sns.lineplot(data=chinese_users, x='Year', y='Cellular Subscription', label='移
动互联网订阅每一百人比例', lw=3)
    sns.lineplot(data=chinese_users, x='Year', y='Broadband Subscription', label='
宽带每一百人订阅比例', lw=3)
    plt.legend(loc='upper left')
    plt.savefig(svgSavePath + '中国互联网用户的数量（单位：千万人）、占人口的比例、移动互联网
订阅每一百人比例、宽带每一百人订阅比例.svg')
```

```
    plt.savefig(pngSavePath + '中国互联网用户的数量（单位：千万人）、占人口的比例、移动互联网订阅每一百人比例、宽带每一百人订阅比例.png')
    plt.show()


    # 绘制各项指标的增长率图
    set_seaborn_properties()
    chinese_users.loc[:, 'increase of No. of Internet Users'] = 0
    chinese_users.loc[:, 'increase of Internet Users(%)'] = 0
    chinese_users.loc[:, 'increase of Cellular Subscription'] = 0
    chinese_users.loc[:, 'increase of Broadband Subscription'] = 0
    rows = len(chinese_users.index)
    for i in range(rows - 1):
        chinese_users.loc[:, 'increase of No. of Internet Users'].iloc[i + 1] = 0 if chinese_users.iloc[i]['No. of Internet Users'] == 0 else (chinese_users.iloc[i + 1].loc['No. of Internet Users'] - chinese_users.iloc[i]['No. of Internet Users']) / chinese_users.iloc[i]['No. of Internet Users']
        chinese_users.loc[:, 'increase of Internet Users(%)'].iloc[i + 1] = 0 if chinese_users.iloc[i]['Internet Users(%)'] == 0 else (chinese_users.iloc[i + 1]['Internet Users(%)'] - chinese_users.iloc[i]['Internet Users(%)']) / chinese_users.iloc[i]['Internet Users(%)']
        chinese_users.loc[:, 'increase of Cellular Subscription'].iloc[i + 1] = 0 if chinese_users.iloc[i]['Cellular Subscription'] == 0 else (chinese_users.iloc[i + 1]['Cellular Subscription'] - chinese_users.iloc[i]['Cellular Subscription']) / chinese_users.iloc[i]['Cellular Subscription']
        chinese_users.loc[:, 'increase of Broadband Subscription'].iloc[i + 1] = 0 if chinese_users.iloc[i]['Broadband Subscription'] == 0 else (chinese_users.iloc[i + 1]['Broadband Subscription'] - chinese_users.iloc[i]['Broadband Subscription']) / chinese_users.iloc[i]['Broadband Subscription']
    plt.title('中国互联网用户的数量（单位：千万人）、占人口的比例、移动互联网订阅每一百人比例、宽带每一百人订阅比例的增长率')
    plt.xlabel('年份')
    plt.ylabel('数值')
    sns.lineplot(data=chinese_users, x='Year', y='increase of No. of Internet Users', lw=4,
                 label='数量（单位：千万人）增长率')
    sns.lineplot(data=chinese_users, x='Year', y='increase of Internet Users(%)', lw=4,
                 label='占人口的比例')
    sns.lineplot(data=chinese_users, x='Year', y='increase of Cellular Subscription', lw=4,
                 label='移动互联网订阅每一百人比例增长率')
    sns.lineplot(data=chinese_users, x='Year', y='increase of Broadband Subscription', lw=4,
                 label='宽带每一百人订阅比例增长率')
    plt.legend(loc='upper left')
    plt.savefig(svgSavePath + '中国互联网用户的数量（单位：千万人）、占人口的比例、移动互联网订阅每一百人比例、宽带每一百人订阅比例的增长率.svg')
```

```
    plt.savefig(pngSavePath + '中国互联网用户的数量（单位：千万人）、占人口的比例、移动互联网订阅每一百人比例、宽带每一百人订阅比例的增长率.png')
    plt.show()

    # 利用多元线性回归预测中国互联网到2050年的总用户数
    # 拟合：
    set_seaborn_properties()
    plt.title('对1980到2020年中国互联网总用户数的拟合')
    sns.scatterplot(data=chinese_users, x='Year', y='No. of Internet Users')
    poly_reg = PolynomialFeatures(degree=3)
    x = chinese_users[['Year']]
    x_m = poly_reg.fit_transform(x)

    model_2 = linear_model.LinearRegression()
    model_2.fit(x_m, chinese_users[['No. of Internet Users']])
    data = pd.DataFrame({'x': x['Year'], 'pred_y': [x[0] for x in model_2.predict(x_m)]})
    plt.xlabel('年份')
    plt.ylabel('人数（单位：千万人）')
    sns.lineplot(data=data, x='x', y='pred_y')
    plt.savefig(svgSavePath + '对1980到2020年中国互联网总用户数的拟合.svg')
    plt.savefig(pngSavePath + '对1980到2020年中国互联网总用户数的拟合.png')
    plt.show()

    # 预测：
    set_seaborn_properties()
    plt.title('到2030年中国互联网总用户数的预测')
    plt.xlabel('年份')
    plt.ylabel('人数（单位：千万人）')
    pred_x = pd.DataFrame(np.arange(1980, 2031), columns=['Year'])
    pred_x_m = poly_reg.fit_transform(pred_x)
    plt.plot(pred_x, model_2.predict(pred_x_m))
    plt.savefig(svgSavePath + '到2030年中国互联网总用户数的预测.svg')
    plt.savefig(pngSavePath + '到2030年中国互联网总用户数的预测.png')
    plt.show()


if __name__ == '__main__':
    # 检查并创建目录
    savePath = './img/Windows/' if sys.platform == 'win32' else './img/MacOS/' if sys.platform == 'darwin' else './img/Linux/'
    pngSavePath = savePath + 'png/'
    svgSavePath = savePath + 'svg/'
    if not os.path.exists(pngSavePath):
```

```
        os.makedirs(pngSavePath)
    if not os.path.exists(svgSavePath):
        os.makedirs(svgSavePath)
    # 读取文件，获取全球互联网用户信息
    global_users = pd.read_csv('Final.csv', delimiter=',', usecols=range(1, 8))  # 由于第一列的列名未知，所以不使用第一列
    # 对全球用户进行分析：
    global_internet_users_analysis()
    entities_2020_internet_users_percentage_pie_bar()
    entities_2020_internet_users_percentage_distribution_histogram()
    entities_2020_internet_users_percentage_distribution_scatter()
    draw_internet_users_percentage_annual_top_3_wordcloud()

    # 通过切片获取中国互联网用户信息
    chinese_users = global_users.loc[global_users['Entity'] == 'China']
    chinese_users_analysis()
```