**《数据可视化》课程设计报告**

设计题目：居民收入和消费支出情况分析

专 业： 数据科学与大数据技术

指导教师： 张军 宋沛 邓涔

学 号： 202012769

姓 名： 姚远

同组人员： 陈晨 张玉豪

计算机与数据科学学院

2023 年 6月1 日

《数据可视化课程设计》评价量表

**任务描述：**通过课程设计把所学的理论知识应用于实际问题的解决，基于课程内容和数据分析方法撰写出规范的数据可视化分析文档，锻炼学生的分析解决实际问题的能力，强化实践意识、提高实际动手能力和创新能力，为以后文档撰写的标准化打基础。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 典范 | 合格 | 有待改善 | 评价 |
| 设计、开发解决方案能力70% | * 选题合理、内容具体 * 有明确的探索性分析 * 融合多种数据分析方法 * 可视化选择准确 * 有明确结论 * 报告思路清晰 * 可视化大屏美观 * 格式规范 * 掌握不同类型的数据的可视化方法 * 可以优化可视化展示 * 体现学生的问题分析能力 | * 选题有一定的意义 * 对数据进行分析 * 有体现数据分析方法 * 有基本的数据可视化 * 对数据提出结论 * 报告整体上思路较清晰，但内容存在一些问题。 * 格式相对规范 * 数据的可视化方法相对单一 * 可视化部分有体现部分优化 * 主动的问题分析能力有部分体现 | * 选题无意义 * 不符合要求 * 报告内容存在较严重的错误 * 内容过于简单 * 出现雷同的情况 * 没有可视化大屏 * 格式不规范 * 数据的可视化方法单一 * 可视化没有优化 * 无主动的问题分析能力 |  |
| 个人与团队能力  30% | * 团队有明确分工 * 每人角色定位合理 * 每人能胜任角色任务且出色完成 * 个人总结部分逻辑鲜明 | * 团队分工不明 * 成员职责不分 * 个人文档和总结不能准确描述自己的工作 | * 团队管理混乱，分工不明 * 成员职责不清 * 论文整体较差或个人完成部分与要求差别较大 |  |
| 评分： |  | | |  |

目录

[1 开发环境及技术 1](#_Toc136533362)

[1.1 开发环境 1](#_Toc136533363)

[1.2 相关技术 1](#_Toc136533364)

[2 数据 2](#_Toc136533365)

[2.1 数据来源及介绍 2](#_Toc136533366)

[2.2 重要指标介绍 3](#_Toc136533367)

[2.2.1 城镇化率 3](#_Toc136533368)

[2.2.2 泰尔指数 3](#_Toc136533369)

[3 数据预处理 5](#_Toc136533370)

[3.1 数据读取 5](#_Toc136533371)

[3.2 重复值处理 6](#_Toc136533372)

[3.3 缺失值处理 6](#_Toc136533373)

[4 可视化实现 9](#_Toc136533374)

[4.1 设计思路 9](#_Toc136533375)

[4.1人均收入与消费支出排行榜 9](#_Toc136533376)

[4.2全国及各地区消费占比 10](#_Toc136533377)

[4.3部分地区城镇人口数及城镇化率 11](#_Toc136533378)

[4.4泰尔指数情况一览 12](#_Toc136533379)

[4.5人均可支配收入情况一览 14](#_Toc136533380)

[4.6全国人口组成情况一览 16](#_Toc136533381)

[5 可视化大屏 18](#_Toc136533382)

[5.1 生成可视化大屏 18](#_Toc136533383)

[5.2 可视化大屏展示 18](#_Toc136533384)

[6 结论及建议 19](#_Toc136533385)

# 1 开发环境及技术

## 1.1 开发环境

开发环境如下表1.1所示:

表1.1 开发环境

|  |  |
| --- | --- |
|  | 开发环境 |
| 操作系统 | Windows 10 |
| 开发工具 | Anaconda 4.10.1 |
| 语言 | Python 3.8.8 |
| 浏览器 | Microsoft Edge |

## 1.2 相关技术

本课题中用到的技术有Pyecharts,Page和html**。**

Pyecharts是一个用于生成Echarts图表的类库，它可以生成独立的网页，也可以在flask,Django中集成使用。Pyecharts的页面组件Page可以将多个图表汇总到一个页面中，方便数据可视化报告的制作。

Page组件是一个容器，可以将多个图表放在一个页面中，支持多种布局方式。Page组件提供了多种布局方式，如Grid、SimplePageLayout、DraggablePageLayout等，同时，还提供了一些常用的配置项，如标题、背景颜色、全局字体等。

HTML是一种标记语言，用于创建网页。在Pyecharts中，可以使用render()方法将图表渲染为HTML文件。

# 2 数据

## 2.1 数据来源及介绍

本项目所选的数据集是来自和鲸社区的数据集，共两个数据集，分别是“居民人均消费支出.txt”和“2000-2019年分地区人均可支配收入、城市和农村可支配收入以及泰尔指数.xlsx”。

部分数据展示分别如下图2.1、图2.2所示：

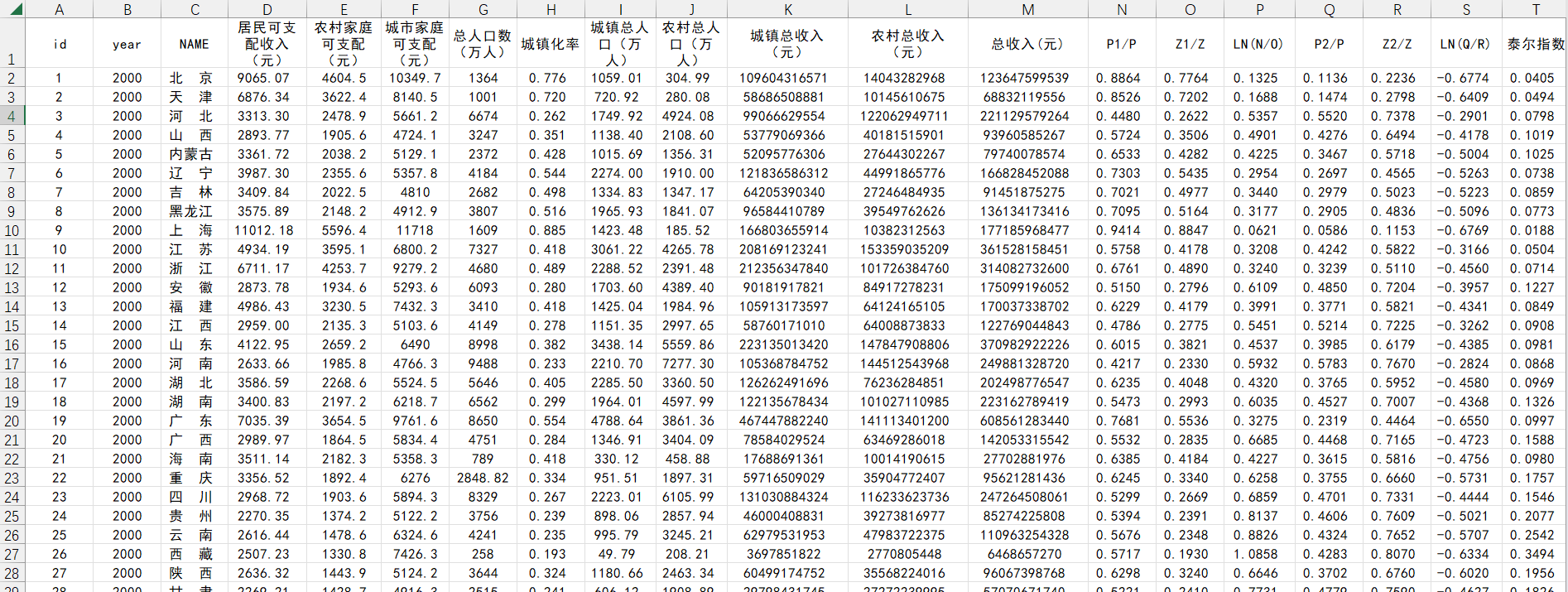


图2.1 泰尔指数数据集



图2.2 人均消费支出数据集

其中“居民人均消费支出.txt”数据集包含了全国及各地区的人均收入、消费支出和具体不同类型消费额度数据；“2000-2019年分地区人均可支配收入、城市和农村可支配收入以及泰尔指数.xlsx”数据集包含了2000年到2019年中国各地区的人均可支配收入、城市和农村可支配收入以及泰尔指数等指标。

## 2.2 重要指标介绍

### 2.2.1 城镇化率

城镇化率是指一个国家、地区或特定地点的城镇人口占总人口的比例。它通常用百分比表示。城镇化率的计算基于居住在城市和乡村的人口之间的区别。城市通常具有更高的人口密度、更多的经济和社会机会以及更发达的基础设施，而乡村则具有更多的农业和自然资源。

城镇化率的提高反映了人们从农村地区迁移到城市地区的趋势。这种趋势通常与工业化、经济发展和社会变革有关。城镇化率的增长可以导致城市人口的增加，城市基础设施的建设和扩张，以及农村地区的人口减少和农业活动的减少。

城镇化率的高低对一个国家或地区的发展具有重要的影响。高度城镇化的国家通常拥有更多的就业机会、更丰富的经济活动和更高的生活水平。然而，城镇化也可能带来一些挑战，例如城市贫困、环境问题和基础设施压力。

城镇化率是衡量一个地区发展和人口分布的重要指标，政府和研究机构经常使用它来评估和监测城市化的进程和影响。

### 2.2.2 泰尔指数

泰尔指数的另一个名称是泰尔熵标准 (Theil’s entropy measure)。这一指数经常被使用。泰尔熵标准是由泰尔 (Theil,1967)利用信息理论中的熵概念来计算收入不平等而得名。

泰尔指数（Theil Index）是一种常用的不平等度量指标，用于衡量某个领域内的收入或财富分布的不平等程度。它可以用来评估一个国家、地区或群体内部的经济不平等状况。

泰尔指数的计算方法涉及个体收入或财富的相对比例。它基于信息熵的概念，通过将整体的不平等性与个体之间的差异相结合，来度量不平等的程度。

泰尔指数的取值范围通常在0到1之间，其中0表示完全平等，1表示最大的不平等。如果泰尔指数为0.5，表示一半的总收入或财富集中在少数一半的人手中。

计算泰尔指数的方法有多种，其中一种常用的方式是通过将个体的收入或财富按照大小排序，然后计算相对差异的平均值。具体的计算公式可能会根据研究领域和具体的数据集而有所不同。

泰尔指数在经济学、社会学和政治学等领域得到广泛应用，可以用于比较不同国家或地区的收入或财富分配差异，以及跟踪不平等的变化趋势。它提供了一个定量的指标，用于理解和评估不平等问题，并帮助政策制定者采取相应的措施来解决或减轻不平等。

# 3 数据预处理

## 3.1 数据读取

由于本次开发所应用的数据集分别有TXT与XLSX两种格式的文件，因此，读取数据时分别调用pandas函数库中的read\_csv()与read\_excel()方法对两个数据集进行读取。代码及结果图分别如下图3.1、图3.2所示：

data\_1 = pd.read\_excel("2000-2019年分地区人均可支配收入、城市和农村可支配收入以及泰尔指数.xlsx")

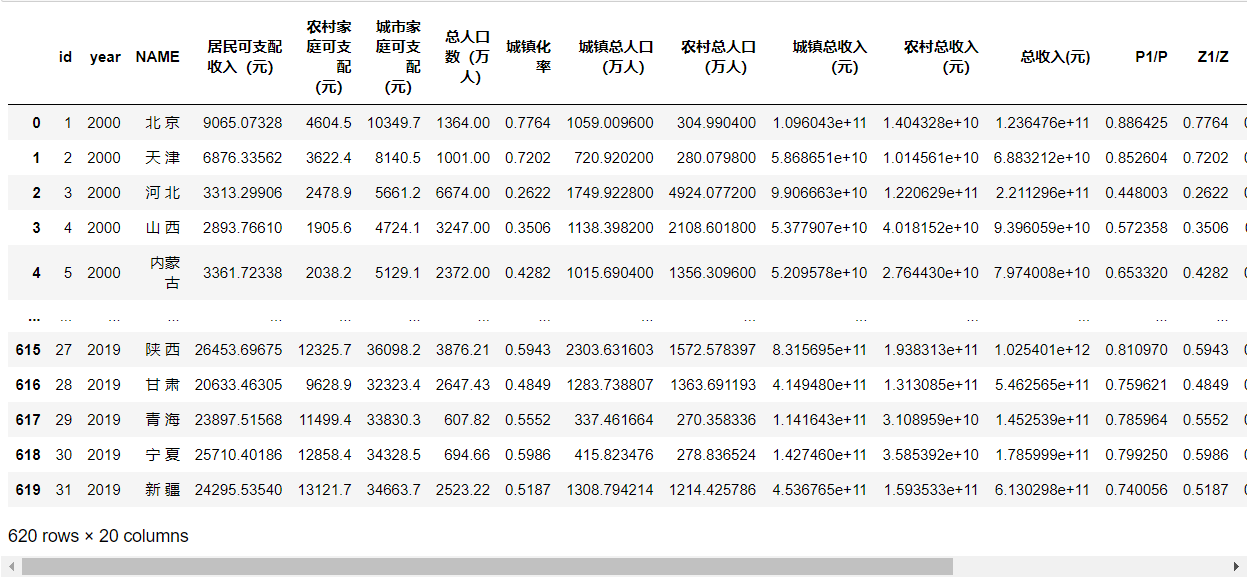


图3.1 读取泰尔指数数据

data\_2 = pd.read\_csv("居民人均消费支出.txt",sep=' ')



图3.2 读取人均消费支出数据

## 3.2 重复值处理

重复值是指样本数据中某个或某些数据记录完全相同，主要是由于人工录入、机械故障等导致部分数据重复录入。重复值主要有两种处理方式，即删除重复值和保留重复值，其中删除重复值是比较常见的方式，其目的在于保留唯一的数据记录。需要说明的是，在分析演变规律、样本不均衡处理、业务规则等场景中，重复值具有一定的使用价值，需保留。

在本次开发的读取数据阶段，同时对数据进行了去重操作，具体代码如下：

data\_1.drop\_duplicates(inplace=True)

data\_2.drop\_duplicates(inplace=True)

执行这一步的操作后，再通过调用pandas中的duplicated()方法，对两个数据集分别进行重复值检测，结果显示不再存在重复值。具体代码及结果图分别如下图3.3、图3.4所示：

data\_1[data\_1.duplicated()]

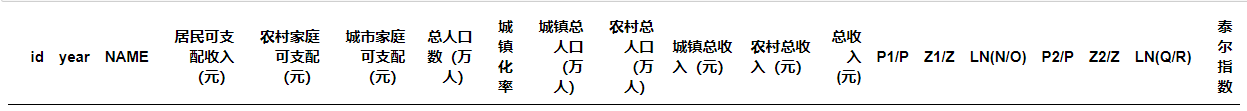


图3.3 检测泰尔指数数据重复值

data\_2[data\_2.duplicated()]



图3.4 检测人均消费支出数据重复值

## 3.3 缺失值处理

缺失值是指样本数据中某个或某些属性的值是不全的，主要是由于机械故障、人为因素等导致部分数据未能收集。若使用存在缺失值的数据进行分析，则会降低预测结果的准确率，需通过合适的方式予以处理。缺失值主要有3种处理方式：删除、填充和插补。

对于本次开发，我们通过使用info()方法分别查看数据集信息的方式查看是否存在缺失值，由结果可以看出数据集不存在缺失值，故不再进行缺失值的处理。具体代码及结果图分别如下图3.5、图3.6所示：

data\_1.info()

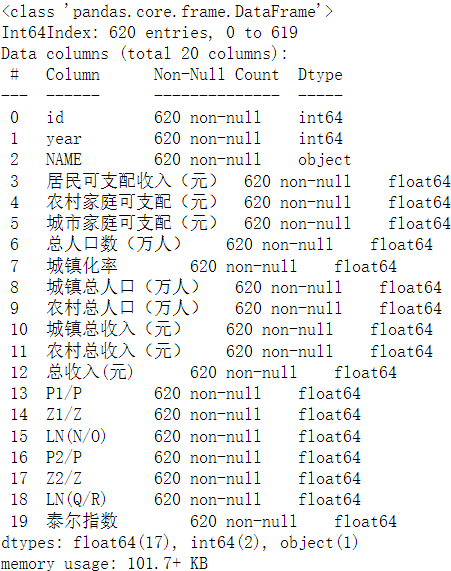


图3.5 泰尔指数数据集信息

data\_2.info()

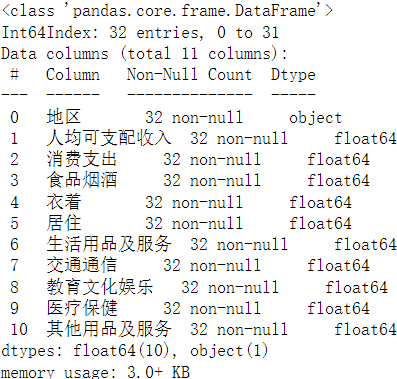


图3.6 人均消费支出数据集信息

# 4 可视化实现

## 4.1 设计思路

此次开发的设计思路如下图4.1所示：

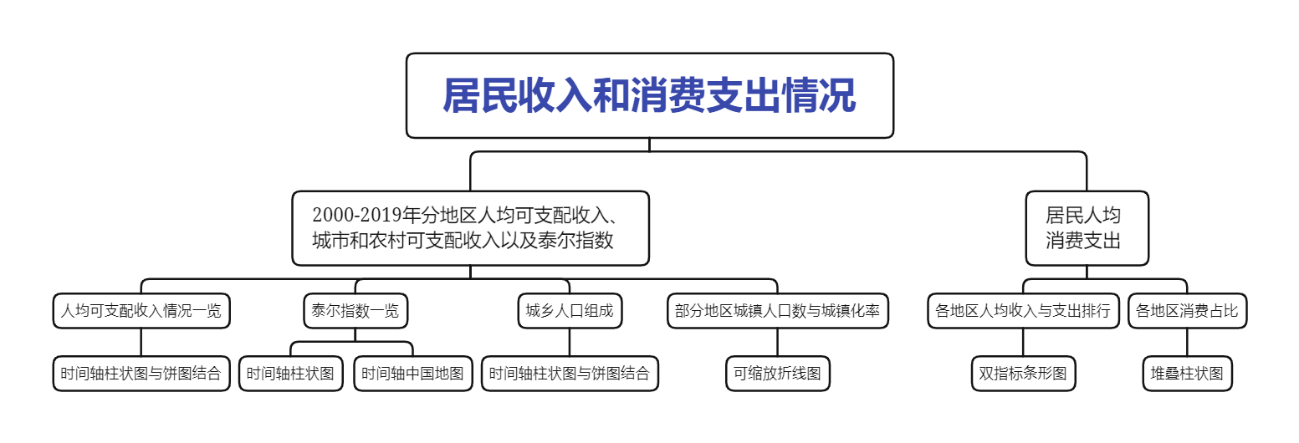


图4.1 设计思路

## 4.1人均收入与消费支出排行榜

人均收入与消费支出排行榜的绘制使用了Pyecharts库来创建一个条形图，展示了全国各地区的人均可支配收入和消费支出情况对比。

首先，我们准备了需要绘制的数据，包括地区、人均可支配收入和消费支出。

接下来，我们创建了一个Bar实例，并通过add\_xaxis()方法设置横坐标为地区列表，通过add\_yaxis()方法添加了两个系列，分别是人均可支配收入和消费支出，对应的纵坐标分别是收入和支出的数值列表。

然后，我们通过reversal\_axis()方法将图表的横纵坐标进行反转，使地区名称在纵轴显示。通过set\_series\_opts()方法设置了标签选项，将标签位置设置在右侧。

最后，我们通过set\_global\_opts()方法设置了图表的全局选项，包括标题、横轴名称、纵轴名称等，然后使用render\_notebook()方法将图表渲染在Notebook中，以便直接在Notebook中显示图表。如图4.2所示：

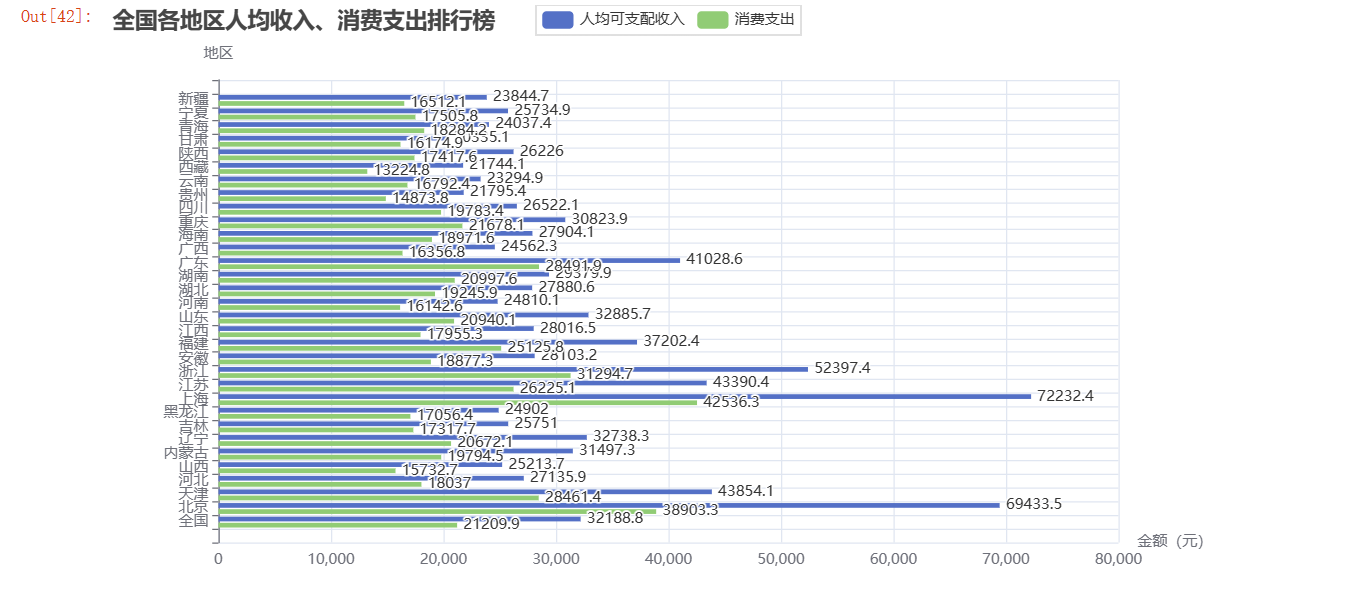


图4.2 人均收入与消费支出排行榜

由图可知，北京、上海、浙江等发达地区在人均可支配收入和消费支出方面表现出较高的水平。这表明这些地区的经济相对较为发达，人们的收入水平相对较高，并且有较高的消费水平。

相比之下，一些经济相对较为欠发达的地区，如贵州、西藏等地，人均可支配收入和消费支出水平较低。这可能反映了这些地区的经济发展相对滞后，人们的收入水平较低，消费水平较低。

## 4.2全国及各地区消费占比

全国及各地区消费占比的绘制使用了Pyecharts库创建了一个堆叠条形图，用于展示全国及各地区不同消费类别的占比情况。

首先，通过导入必要的模块，创建了一个Bar实例。

然后，使用add\_xaxis方法设置横坐标为地区列表，使用add\_yaxis方法添加了多个系列，分别代表不同消费类别的占比情况。每个系列的纵坐标数据都对应了数据集中相应类别的占比值。

接下来，使用set\_global\_opts方法设置了图表的全局选项，包括标题、图例等。标题设置为"全国及各地区消费占比"，并通过pos\_left参数调整标题位置。图例设置在右侧中部，以垂直方式显示。

最后，使用set\_series\_opts方法对系列进行设置，其中label\_opts参数设置为不显示数据标签。使用render\_notebook方法将图表渲染在Notebook中，以便直接在Notebook中显示图表。如图4.3所示：

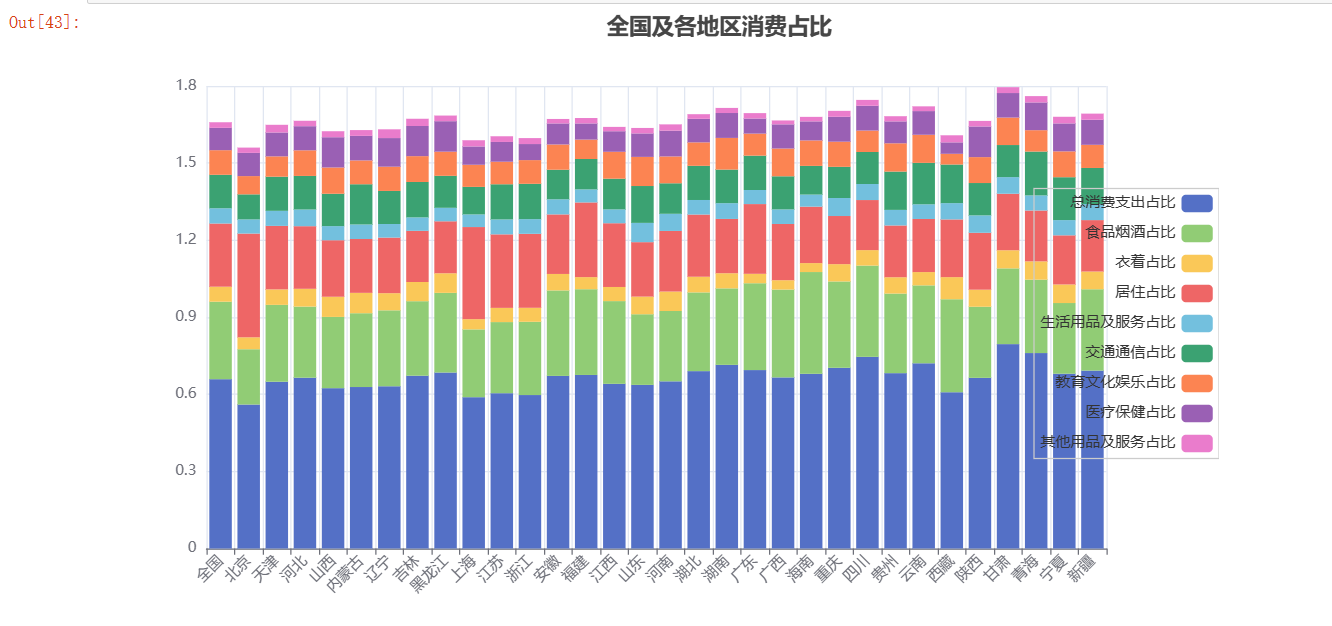


图4.3 全国及各地区消费占比

通过这个堆叠柱状图，我们可以直观地比较各地区不同消费类别的占比情况。通过观察图表，可以了解各类别在消费中的相对重要程度，并对不同地区的消费结构进行比较和分析，得出结论：

交通通信占比和居住占比在各地区的消费中较大。这意味着人们在日常消费中对交通和通信以及居住方面的支出相对较高。这可能是由于交通和通信的发展水平以及居住成本在不同地区的差异导致的。

教育文化娱乐占比在一些地区相对较低。这可能是因为在一些地区的消费结构中，教育、文化和娱乐方面的支出相对较少，或者人们在这些方面的消费需求相对较低。

## 4.3部分地区城镇人口数及城镇化率

这段代码定义了两个函数。首先，cal\_lis1()函数根据给定的地区名称，从数据集中提取相应地区的城镇化率和城镇总人口数据，并返回处理后的数据。同时，该函数还计算了给定地区的城镇总人口的最大值。

接下来，plot\_mul\_line()函数根据传入的参数绘制了一个复合折线图。该图包含两个折线图，分别表示城镇人口数和城镇化率随时间的变化情况。函数根据传入的地区名称、标题、时间数据、城镇化率数据以及城镇人口数据进行绘制。图表的样式和配置通过使用Pyecharts库中的相关选项进行设置。

在主程序中，我们从“全国各地区人均收入、消费支出排行榜”图中选择人均收入与消费支出TOP3的城市进行数据提取和绘图，并获取了它们的城镇化率和城镇总人口数据，以及最大人口数。然后，通过调用plot\_mul\_line()函数，并传入相关数据，绘制了城镇人口数和城镇化率的关系图表。如图4.4所示：

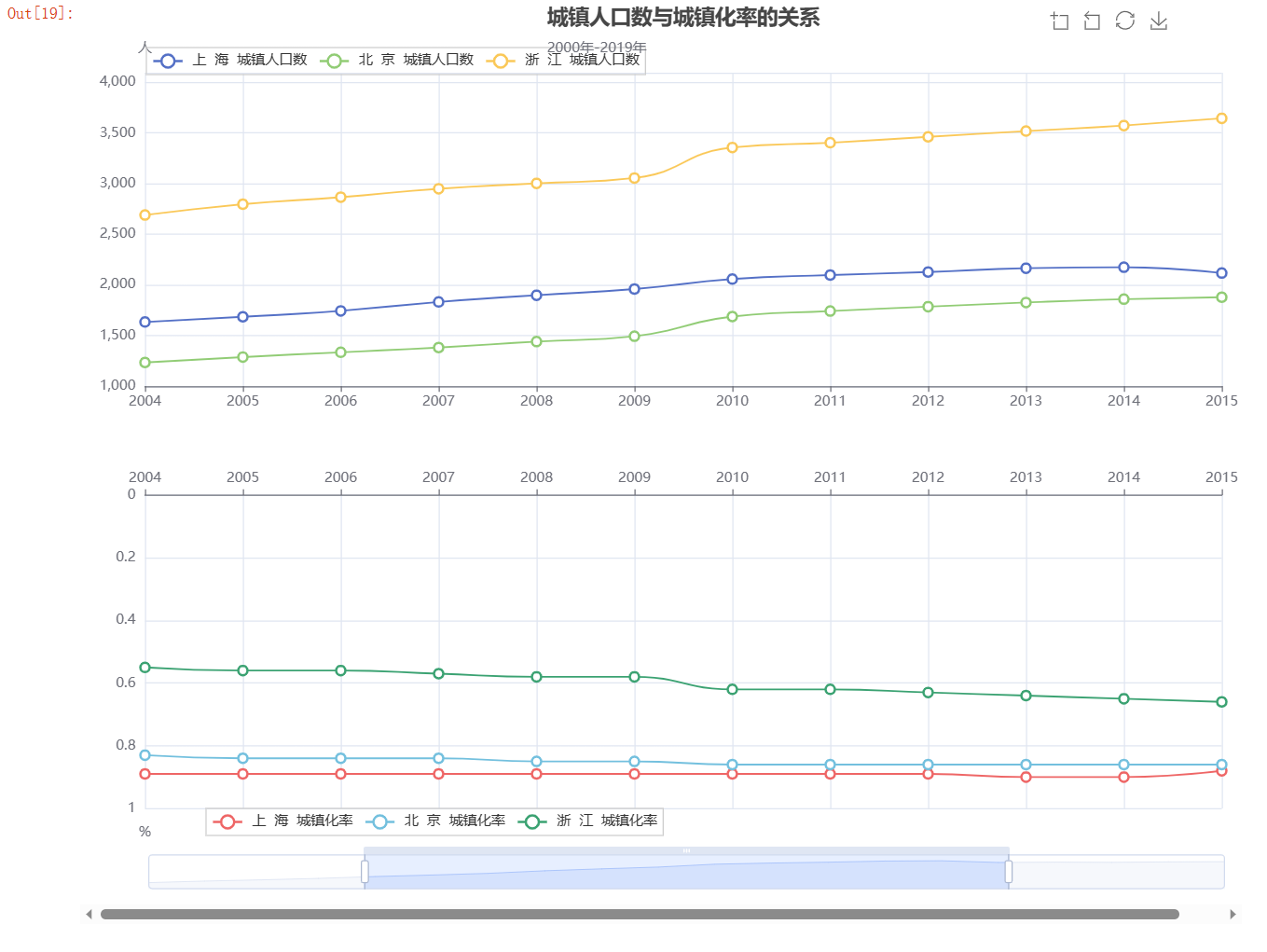


图4.4 部分地区城镇人口数及城镇化率

通过比较上图的上下两部分可知，虽然城镇人口有所提升，但是城镇化率的提升并没有与城镇人口的保持一致，而是逐渐趋于平稳。三个城市中城镇化率最高的是上海，对比上海的城镇人口增长，两者的增速都近乎平稳。浙江的城镇人口增速是三个城市中最明显的，但是其城镇化水平却是三个城市中最低的。

## 4.4泰尔指数情况一览

4.4.1 柱状图

泰尔指数是衡量收入分配不均衡程度的指标，指数值越高表示收入分配越不均衡。这段代码使用了Pyecharts库来绘制时间轴图表，并展示了不同年份的柱状图。首先，通过调用get\_year\_overlap\_chart()函数，生成了一个柱状图（Bar）对象，用于表示指定年份的泰尔指数情况。

然后，创建了一个时间轴对象，即timeline1，用于按照年份的顺序展示不同年份的柱状图。通过一个循环，将每个年份对应的柱状图添加到时间轴中，使用timeline1.add()实现。

最后，通过设置时间轴的相关参数，如自动播放和播放间隔，使用timeline1.add\_schema()。

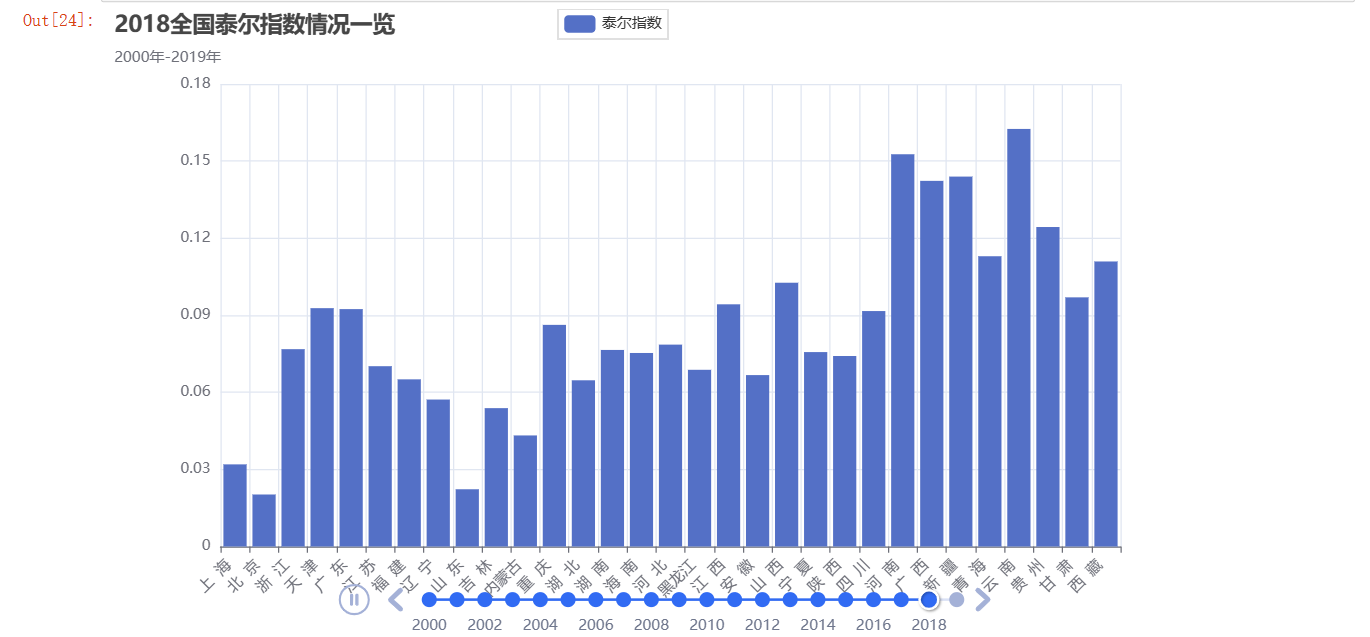
最终，通过调用timeline1.render\_notebook()方法，将时间轴图表渲染为Jupyter Notebook中的可视化结果。如图4.5所示：  


图4.5 泰尔指数情况一览柱状图

由图可知，不同地区之间的泰尔指数存在差异。一些地区如新疆，贵州在多个年份都呈现较高的泰尔指数，说明这些地区的收入分配问题比较严重。而另一些地区如北京，上海的泰尔指数相对较低，说明这些地区的收入分配相对较为均衡。

综上所述，通过时间轴图表，我们可以直观地观察到全国各地区在收入分配不均衡方面的变化趋势，并比较不同地区的泰尔指数情况。这有助于我们了解和关注收入分配的公平性和社会经济发展的平衡性。

4.4.2地图

这段代码首先创建了一个空列表name\_list，用于存储统一格式后的地区名称。然后使用循环遍历原始地区名称列表names，对每个地区名称进行处理，将不规范的名称统一为规范的格式，并添加到name\_list中。例如，将"上海"统一为"上海市"，将"重庆"统一为"重庆市"，将"天津"统一为"天津市"等。

重新定义了data\_ind字典，并将格式化后的数据存储到其中。每个年份对应一个子字典，其中包含每个地区的名称和对应的泰尔指数值。

修改了format\_data函数，将地区名称与对应的泰尔指数值一起存储，并使用统一格式的地区名称。

创建了一个新的时间轴对象timeline2，并使用循环将每个年份的地图图表添加到时间轴中。

调用get\_year\_overlap\_chart()函数获取每个年份对应的地图（Map）对象，表示该年份的泰尔指数情况。

在地图对象中，使用add()方法添加地图数据，设置标题、视觉映射配置和提示框等全局选项。

最后，通过调用timeline2.render\_notebook()方法，将时间轴图表渲染为Jupyter Notebook中的可视化结果。如图4.6所示：

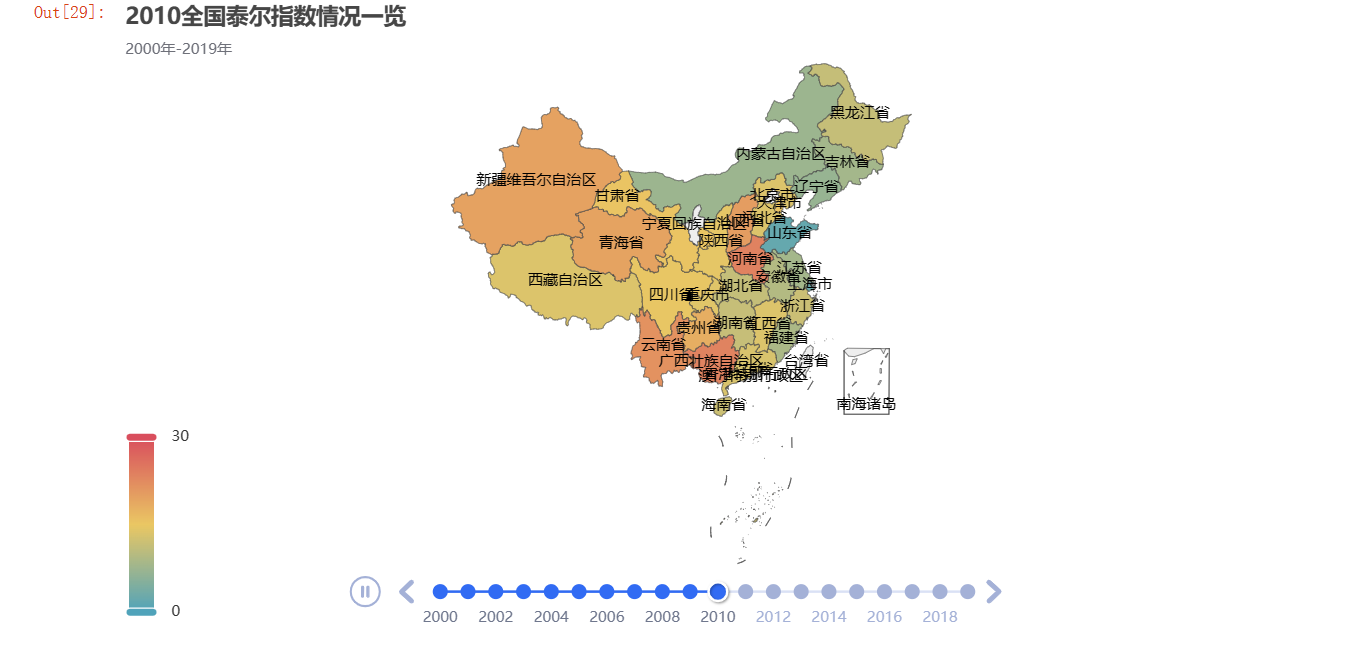


图4.6 泰尔指数情况一览地图

根据结果图，通过时间轴的方式，可以逐年查看泰尔指数的变化趋势，并对不同地区之间的差异进行比较。

## 4.5人均可支配收入情况一览

在该主题的代码实现中定义了一个名为get\_year\_overlap\_chart的函数，用于生成可支配收入情况的时间轴图表。函数内部的操作包括：

1.创建了一个柱状图（Bar）对象bar，用于展示不同年份的居民、城市和农村可支配收入的情况。使用add\_yaxis()方法分别添加居民、城市和农村可支配收入的数据，其中total\_data["dataINCOME"][year]表示居民可支配收入，total\_data["dataCITY"][year]表示城市可支配收入，total\_data["dataCOUNTRY"][year]表示农村可支配收入。

使用set\_global\_opts()方法设置全局选项，包括标题、提示框和横轴标签的显示设置。

2.创建了一个饼图（Pie）对象pie，用于展示城市和农村可支配收入的占比情况。使用add()方法添加饼图的数据，其中城市和农村的可支配收入占比由total\_data["dataCITY"]["{}sum".format(year)]和total\_data["dataCOUNTRY"]["{}sum".format(year)]给出。

使用set\_series\_opts()方法设置饼图的提示框选项。

最后，使用.overlap()方法将柱状图和饼图叠加在一起，并返回叠加后的图表对象。

接下来，代码创建了一个新的时间轴对象timeline3，通过循环遍历每个年份，并调用get\_year\_overlap\_chart(year=y)函数获取对应年份的图表。将每个年份的图表添加到时间轴中，并使用add\_schema()方法设置自动播放和播放间隔。

最后，调用timeline3.render\_notebook()方法将时间轴图表渲染为Jupyter Notebook中的可视化结果。如图4.7所示：

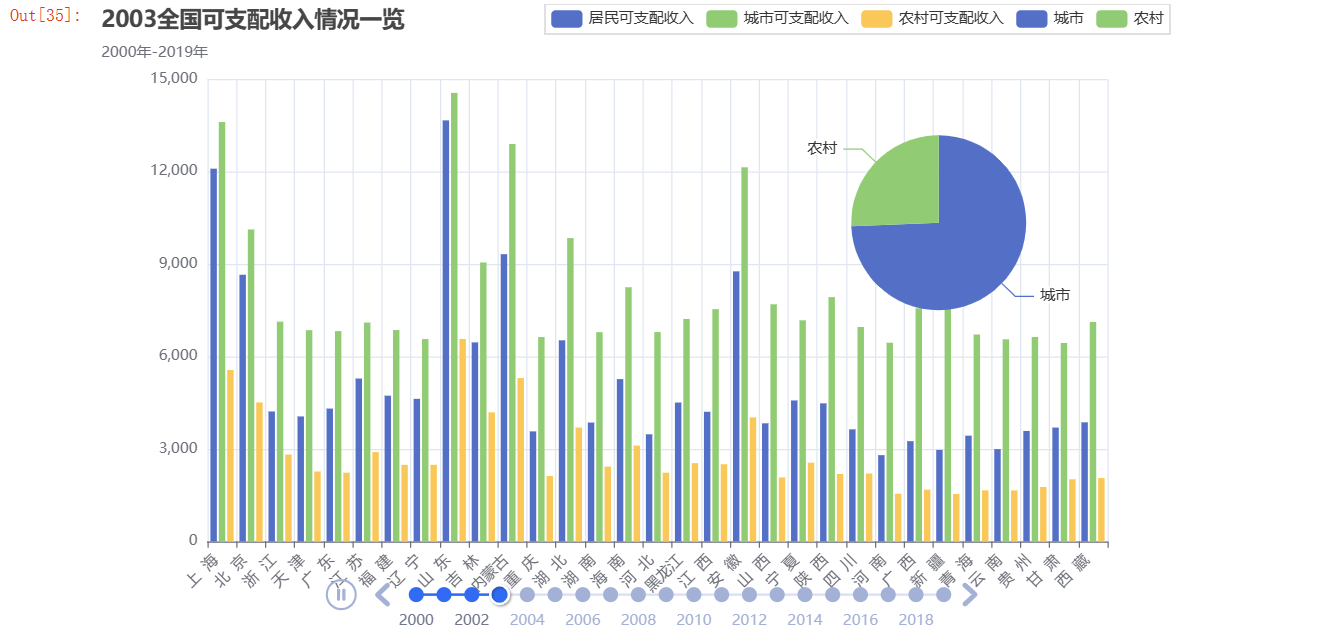


图4.7 人均可支配收入情况一览

该图表的目的是展示不同年份的居民、城市和农村可支配收入情况，并通过柱状图和饼图的叠加来呈现可支配收入的占比情况。时间轴图表可以逐年观察可支配收入的变化趋势，并对不同地区和来源的收入进行比较和分析。通过观察该图表可以得出以下结论：

上海、山东和内蒙古地区的可支配收入较大，表明这些地区的居民、城市和农村收入水平相对较高。

随着时间的推移，全国的可支配收入总体呈现增长趋势，但不同地区和来源的收入增长速度可能存在差异。

饼图显示了城市和农村可支配收入的占比情况。根据图表显示的数据，可以比较城市和农村的收入占比变化，观察城乡收入差距的演变趋势。

通过时间轴的功能，可以逐年观察可支配收入的变化趋势，了解不同地区和来源的收入情况，并进行比较和分析。

## 4.6全国人口组成情况一览

这段代码使用了Python的数据可视化库，主要利用了Bar（柱状图）和Pie（饼图）两个图表类型来绘制全国人口组成情况一览图。具体绘图过程如下：

创建Bar（柱状图）对象，并设置x轴数据为地区名称，分别添加总人口、城镇人口和农村人口的数据作为三个系列。通过设置series\_name和y\_axis来指定系列名称和对应的数据。设置图表的全局选项，包括标题、提示框和x轴标签等。

创建Pie（饼图）对象，并设置系列名称为人口组成，添加城镇人口和农村人口的数据作为两个饼图的数据项。通过设置center和radius来调整饼图的位置和大小。设置饼图的提示框选项，通过tooltip\_opts指定是否显示提示框和触发方式。

使用bar.overlap(pie)将柱状图和饼图叠加在一起，形成最终的图表。创建Timeline（时间轴）对象，并循环添加每个年份的图表，使用get\_year\_overlap\_chart函数生成图表对象。设置时间轴的全局选项，包括自动播放、播放间隔等。最后，使用render\_notebook()将图表渲染并在Notebook中显示。如图4.8所示：

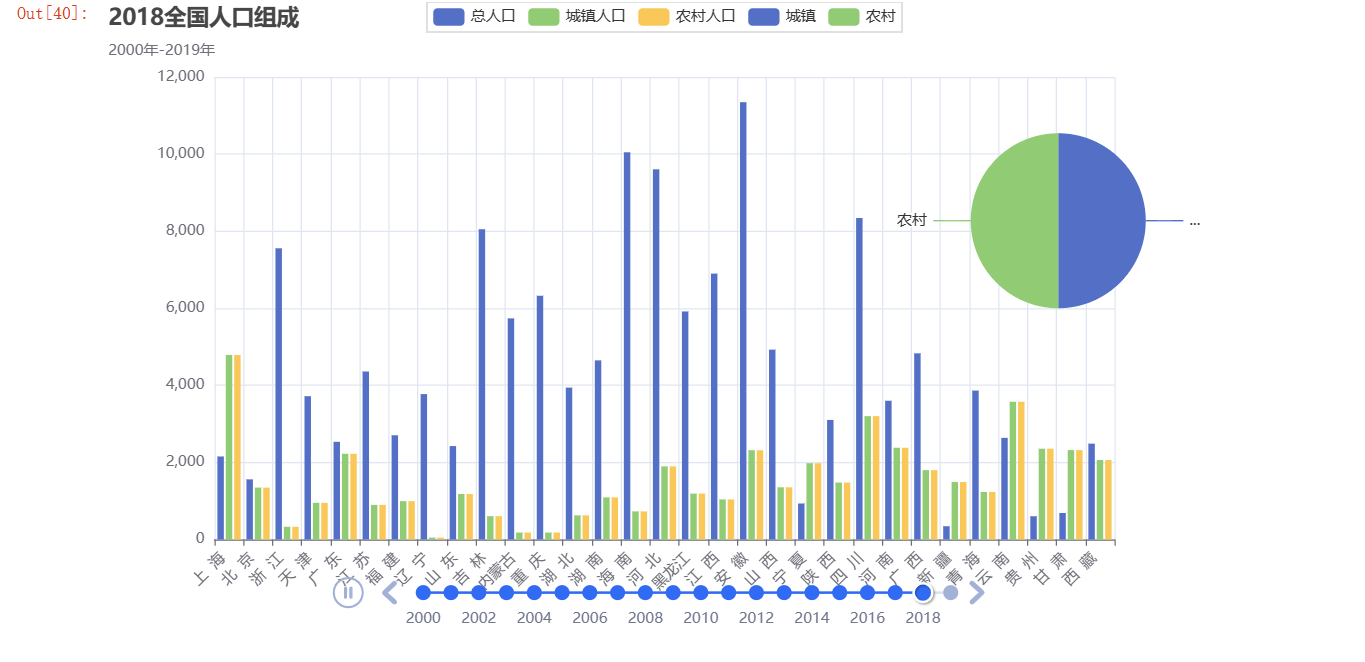


图4.8 全国人口组成情况一览

该图表的目的是同时展示不同年份的全国人口组成情况，通过柱状图和饼图的叠加来呈现人口的数量和比例关系。时间轴的使用使得可以逐年观察人口组成的变化趋势，并进行比较和分析。通过观察该图表可以得出以下结论：

全国的总人口在2000年至2019年期间呈现增长趋势，表示人口总量逐渐增加。随着时间的推移，城镇人口和农村人口的比例可能发生变化。通过饼图可以比较不同年份城镇和农村人口的占比情况，观察城乡人口结构的演变趋势。

柱状图显示了每个年份的总人口、城镇人口和农村人口的数量，可以通过柱状图的高度比较不同年份之间人口的变化情况。可以看出安徽，河北的总居住人口一直居高不下，而新疆，吉林等地区的总居住人口则相对较少且变化较大。

# 5 可视化大屏

## 5.1 生成可视化大屏

使用Pyecharts制作可视化大屏，可以分为三步：

第一步：利用Pyecharts的Page组件将已经绘制完成的图表整合到一个html文件中；

第二步：打开上一步生成的html文件，自定义拖拽布局，完成后另存为js文件；

第三步：再次利用Page组件通过上述两步所生成的html文件和js文件生成存放最终可视化大屏的html文件。

## 5.2 可视化大屏展示

基于Pyecharts的居民收入和消费支出情况分析的可视化大屏如下图5.1所示:

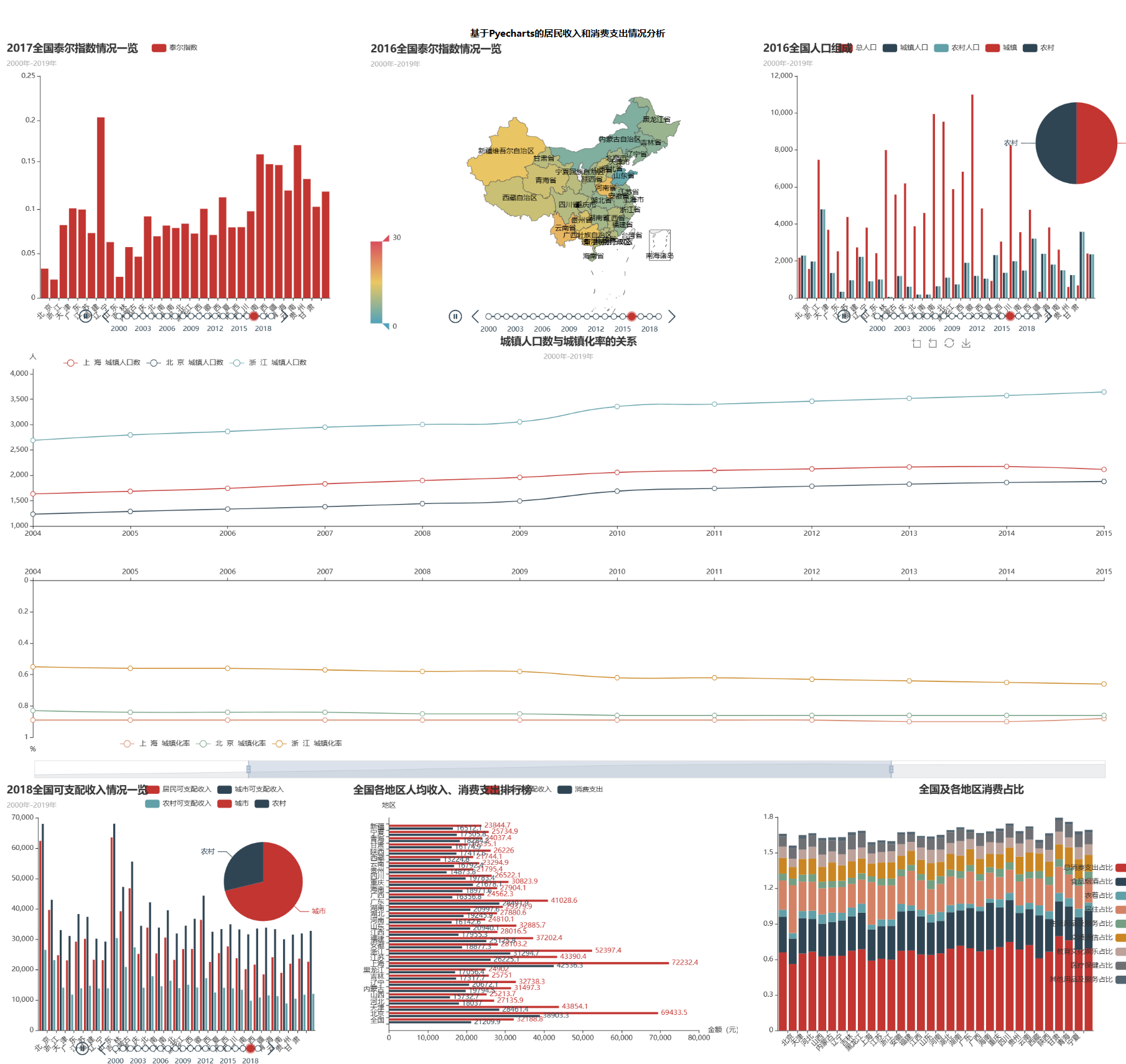


图5.1 可视化大屏展示

# 6 结论及建议

在整个可视化分析及实现过程中，我们可以清晰的了解到随着时间的推进，全国整体及各地区的收入、消费水平都有明显的提高，这也进一步说明我国近年所实施的政策是切实可行且有效的。

但是不同地区的收入与支出是有着较大差异的，这是由于不同地区的消费结构存在较大差异，不同地区收入消费支出都有着各自的偏重点。

为了响应国家号召，更快且更稳定的实现“共同富裕”这一目的，我们提出了一下建议：

第一,通过多种方式提高城镇居民收入，促进居民消费。

收入水平的提高可以促使消费水平进一步提高，因此需通过各种途径提高城镇居民的收入。首先，完善收入分配制度。目前我国居民的收入在GDP中的占比还比较小，其增长速度仍低于经济增长的速度。应该提高居民的劳动工资在收入分配中的比例，企业提高城镇职工的最低工资，政府部门提高个人税收起点，增加可支配收入，进而为消费需求提供信心；其次，完善劳动力市场建设，拓展劳动力就业渠道。政府应加强劳动力市场建设，提供公平的就业机会，改善就业环境，争取更多更好的就业机会，增加居民的收入；最后加强经济宏观调控，调节物价，避免因通货膨胀物价上涨导致居民的可支配收入下降而减少消费指出。

第二，完善城镇居民生活保障，增强居民消费能力。

政府应加大对人民的财政支出，减少人民生活负担，这样就相当于增加了人民的收入，使其消费的能力得到保障，消费质量得到提高。首先，完善财政转移支付制度，提高离退休人员的退休工资，使其消费需求得到满足；其次，加大社会保障力度，提高福利待遇，扩大居民最低生活保障覆盖面，准时足额发放养老金等，增强居民的消费能力；再次，健全医疗保障制度改革，免除居民后顾之忧；最后，增加财政教育投入，这不仅能够增加居民的受教育程度，提高文化知识，提升人力资本，而且减轻了家庭用于子女教育方面的负担，提高了居民家庭的消费信心和消费能力。