

## 国内游客画像模型

### 一、模型概述

本模型聚焦于对来 X 市的国内游客数据进行分析，依据省、直辖市以及其下属城市（区）的层级架构，旨在通过实时数据分析，全方位呈现国内不同地域游客来 X 市的占比情况及分布特征，为 X 市旅游管理部门、旅游企业等深入了解客源市场结构、制定精准营销策略以及优化旅游产品提供直观且准确的数据支撑，助力提升 X 市旅游产业对国内游客的吸引力与服务质量。

### 二、数据来源与整理

**旅游大数据平台：**整合了多个数据源信息的旅游大数据平台发挥着关键作用。它汇聚了来自各大在线旅游平台（OTA）的预订数据（包含酒店预订、景区门票预订等信息，且涉及预订者的出发地等关键内容）、交通票务系统（飞机、火车、长途客车等的购票信息，能明确旅客出发地与目的地，从中筛选出抵达 X 市的国内游客相关数据）以及移动运营商提供的手机信令数据（通过分析游客在 X 市活动期间的信号轨迹，追溯其出发地归属）等，这些海量且多维度的数据构成了分析国内游客来源情况的核心基础。

**景区入园登记系统：**X 市各景区的入园登记记录涵盖了游客的身份证件信息、购票渠道、入园时间等，其中身份证号码前几位可对应游客所在的省、直辖市及城市（区）信息，以此提取出来自不同地域游客进入景区游玩的具体人次情况，从游览环节进一步细化和补充游客来源地数据，对整体分析国内游客地域分布起到重要的验证与完善作用。

**酒店入住登记系统：**酒店作为游客在 X 市的重要停留场所，其入住登记系统记录了住客的详细信息，如常住地址、身份证号码等，通过分析这些字段内容，可以判断住客的来源地，汇总各酒店的数据就能从住宿角度掌握不同地域的游客来 X 市的规模，与景区入园登记数据以及旅游大数据平台的数据相互印证，确保数据的准确性和全面性，更完整地勾勒出国内游客的来源轮廓。

**数据整理流程:** 首先, 将从上述不同渠道收集到的数据汇总到统一的数据仓库中。接着, 运用数据清洗工具和人工核查相结合的方法, 去除重复、错误以及逻辑不清晰的数据记录, 例如纠正因数据录入错误导致的游客来源地标注错误、剔除因系统故障产生的异常票务或登记信息等。然后, 按照省、直辖市以及下属城市(区)的行政划分标准, 通过提取相应字段信息(如身份证号码对应地区代码、预订信息中的出发地等)对游客数据进行分类整理, 将每位游客准确归到所属的地域层级, 为后续的统计分析筑牢基础。

### 三、核心算法与分析逻辑

**省(直辖市)游客占比及总数统计:** 借助数据库管理系统(如 MySQL、Oracle 等)或数据分析软件(如 Python 的数据分析库 pandas 等), 依据整理好的数据中代表游客来源省(直辖市)的字段, 运用分组和计数函数(如 SQL 语句中的 GROUP BY 和 COUNT 函数, 或 pandas 中的 groupby 和 size 方法), 将来自同一省(直辖市)的游客数据归为一组, 统计每组内的游客记录数量, 该数量即为对应省(直辖市)当日来 X 市的游客总数。再通过计算各分组游客数量占全国游客总数的比例, 得出各省(直辖市)来 X 市游客的占比情况, 以此实现对全国省(直辖市)层面游客情况的量化分析。

**下属城市(区)占比统计(点击省(直辖市)后):** 当点击某一省(直辖市)时, 在已筛选出该省(直辖市)游客数据的基础上, 再依据代表游客来源下属城市(区)的字段, 同样运用上述分组和计数方法, 将属于同一城市(区)的游客数据进一步细分并统计数量, 然后计算各城市(区)游客数量占该省(直辖市)游客总数的比例, 得出该省(直辖市)下属各城市(区)来 X 市游客的占比情况, 从而精准呈现省内各城市(区)的游客输出差异。

### 四、模型输出与可视化

**GIS 地图飞线展示(省(直辖市)层面):** 利用地理信息系统(GIS)技术, 以全国地图为底图, 绘制从各省(直辖市)指向 X 市的飞线, 飞线的粗细或颜色深浅可用来表示该省(直辖市)来 X 市游客占比的高低(例如, 占比越高, 飞

线越粗或颜色越深），直观地展示全国不同地域游客流向 X 市的大致分布情况以及占比差异。同时，为地图添加交互功能，当鼠标指针悬停在某一省（直辖市）区域上时，在地图上方或旁边弹出信息框，清晰展示该省（直辖市）的名称以及当日来 X 市游客总数的具体数值。

**柱状图或饼图展示（下属城市（区）层面，点击省（直辖市）后）：**在点击某一省（直辖市）弹出的弹窗中，可采用柱状图或饼图的形式展示该省下属各城市（区）来 X 市游客的占比情况。若使用柱状图，以城市（区）名称为横坐标，各城市（区）游客占比为纵坐标，通过柱子的高低直观对比各城市（区）的游客输出比重差异；若采用饼图，则以各扇形区域代表不同的城市（区），扇形面积大小对应占比高低，同时在每个扇形上标注城市（区）名称及具体占比数值，清晰呈现该省（直辖市）内游客来源的城市（区）结构，方便用户快速查看各城市（区）的占比情况。

## 五、模拟数据示例与可视化展示

假设今日有以下部分省（直辖市）及某省下属部分城市来 X 市旅游的游客数据模拟示例：

省（直辖市）	游客总数（人）
广东省	500
江苏省	300
浙江省	400

对应的各省（直辖市）来 X 市游客占比及飞线可视化展示如下（以简单地图示意）：

（此处应插入一个简单的全国地图示例，有从广东、江苏、浙江等地指向 X 市的飞线，飞线粗细或颜色体现占比情况，鼠标悬停在某省区域上有信息框展示省份名称及游客总数数值示例图）

以下是使用 Python 的 pyecharts 库来生成飞线可视化展示的示例代码:

收起

```
python

from pyecharts import options as optsfrom pyecharts.charts import Map, Geoimport pandas as pd

# 假设数据

data = {

    "省（直辖市）": ["广东省", "江苏省", "浙江省"],

    "游客总数（人）": [500, 300, 400]}

df = pd.DataFrame(data)

# 计算占比

total_visitors = df["游客总数（人）"].sum()

df["占比"] = df["游客总数（人）"] / total_visitors

# 生成地图和飞线

geo = (

    Geo()

    .add_schema(maptype="china")

    .add(

        "",

        [list(z) for z in zip(df["省（直辖市）"], df["占比"])],

        type_="effectScatter",

        color="white",

        symbol_size=10,

        effect_opts=opts.EffectOpts(is_show=False)

    )

    .add(

        "",

        [list(z) for z in zip(df["省（直辖市）"], df["占比"])],

        type_="lines",

        linestyle_opts=opts.LineStyleOpts(width=2, opacity=0.6, curve=0.2),

        label_opts=opts.LabelOpts(is_show=False),

        effect_opts=opts.EffectOpts(symbol="arrow", symbol_size=8, color="blue")

    )

)
```

```

.set_series_opts(label_opts=opts.LabelOpts(is_show=False))

.set_global_opts(
    title_opts=opts.TitleOpts(title="各省（直辖市）到 X 市游客占比及飞线图"),
    visualmap_opts=opts.VisualMapOpts(
        min_=df["占比"].min(),
        max_=df["占比"].max(),
        is_calculable=True,
        range_color=["#313695", "#4575b4", "#74add1", "#abd9e9", "#e0f3f8", "#ffffbf",
        "#fee090", "#fdbe61", "#f46d43", "#d73027"]
    ),
    tooltip_opts=opts.TooltipOpts(
        trigger="item",
        formatter="{b}: {c}"
    )
)
))

# 渲染图表
geo.render("province_to_x_city_flyline.html")

```

在上述代码中：

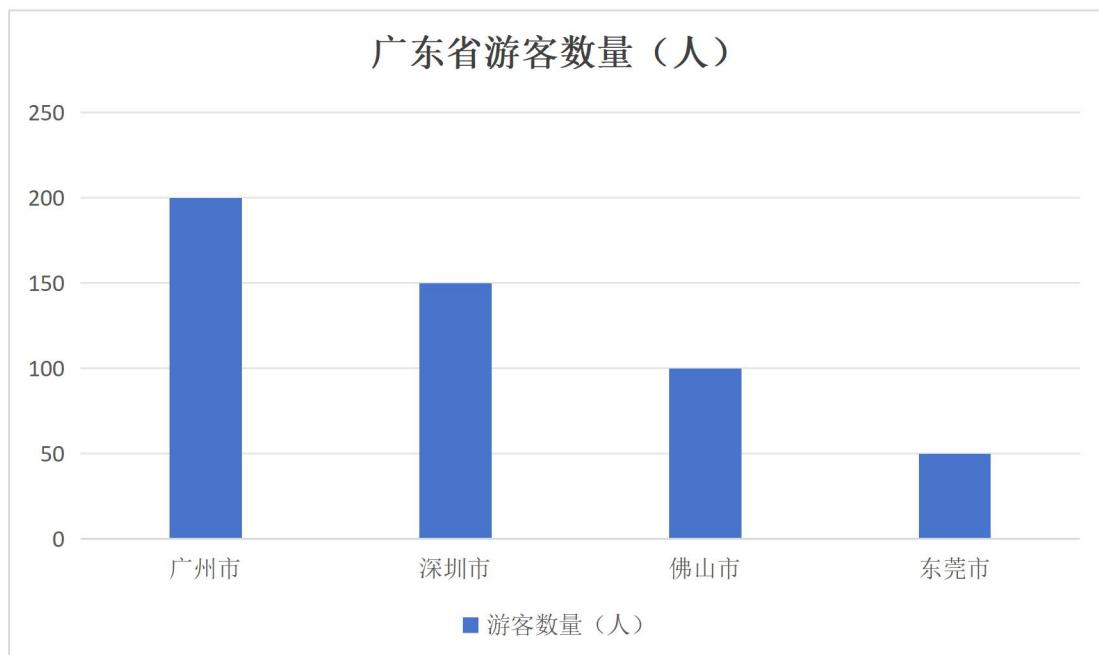
- 首先根据给定的游客总数计算每个省份的占比。
- 然后使用 `Geo` 图表来创建地图和飞线。在 `add_schema` 中设置地图类型为中国地图。
- 第一个 `add` 方法用于添加散点（这里散点主要是为了辅助定位省份位置，设置为白色且无特效显示），第二个 `add` 方法添加飞线，设置了飞线的样式（宽度、透明度、曲线度）、箭头样式和颜色等。
- 通过 `set_global_opts` 设置图表的标题、视觉映射（根据占比设置颜色渐变）和提示框（鼠标悬停时显示省份和占比信息）。
- 最后使用 `render` 方法将图表渲染为 `HTML` 文件，在浏览器中打开该文件即可查看飞线可视化展示。

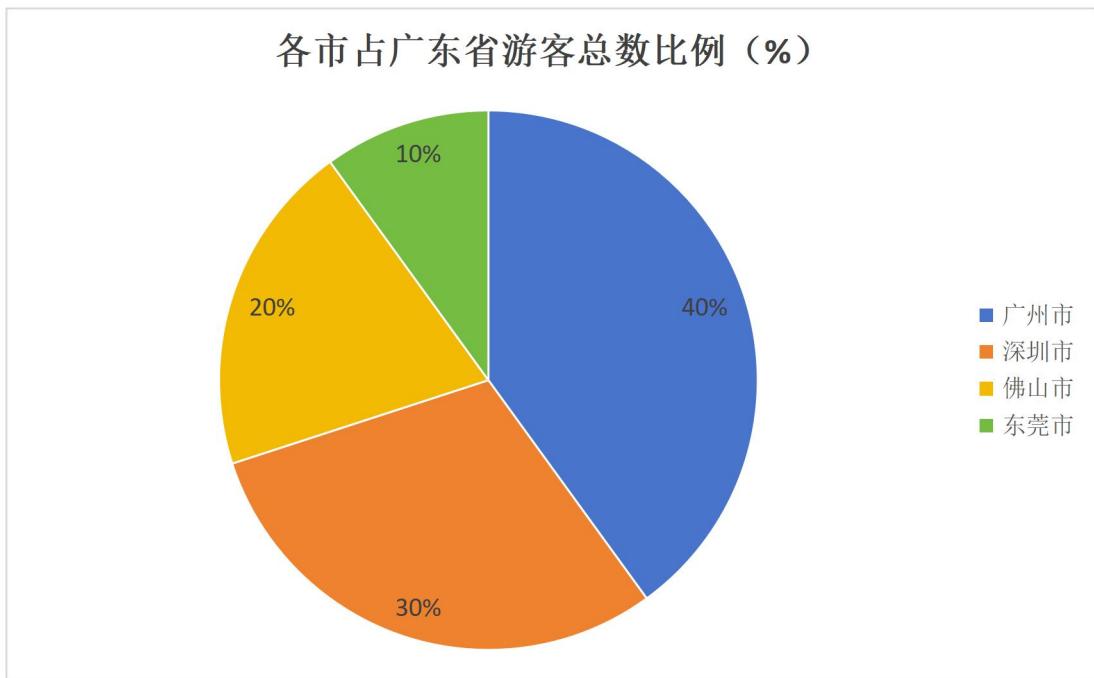
请注意，这只是一个简单的示例，实际应用中可能需要根据具体的数据格式和需求进行调整。如果有更详细的地理坐标信息，还可以进一步优化飞线的起点和终点位置等。

假设广东省下属部分城市来 X 市旅游的游客数据及占比模拟如下：

城市	游客数量 (人)	占广东省游客总数比例 (%)
广州市	200	40
深圳市	150	30
佛山市	100	20
东莞市	50	10

对应的广东省下属城市占比可视化展示（以点击广东省为例，分别展示柱状图和饼图示例）如下：





(此处应插入一个简单的柱状图示例，横坐标为广州市、深圳市、佛山市、东莞市，纵坐标为占比数值，柱子高低体现各城市占比差异，鼠标悬停有交互效果示例图；同时插入一个饼图示例，四个扇形分别对应上述四个城市，扇形面积体现占比大小，每个扇形标注城市名称及占比数值，鼠标悬停有交互效果示例图)

## 六、应用场景与价值

**精准营销与市场拓展：**X 市旅游管理部门和旅游企业可依据模型输出的国内游客画像，制定精准的旅游营销方案。针对游客输出量占比较大的省（直辖市）以及下属城市（区），加大宣传推广力度，如在当地举办旅游推介会、投放针对性广告、与当地旅行社合作等；同时，根据不同地域游客的特点、偏好以及消费习惯，设计差异化的旅游产品和线路，提高营销的精准性和有效性，吸引更多国内游客来 X 市旅游，进一步拓展国内旅游市场。

**旅游资源优化配置：**了解不同地域游客来 X 市的规模和占比情况后，能够合理分配旅游资源。例如，对于来自游客输出大省（直辖市）的游客，相应增加景区接待能力（如增加景区开放时间、增设服务设施等）、调配更多的旅游交通运力、安排更多熟悉对应地域文化和语言的旅游服务人员，确保游客在 X 市能获得良好的旅游体验，提高旅游产业的整体运营效率和服务质量。

**旅游产品优化与创新：**通过深入分析各地区游客的特征差异，有助于 X 市旅游企业和相关机构对现有旅游产品进行优化升级，并开发更具针对性的新产品。比如，了解到某些城市（区）的游客对文化体验类项目有较高兴趣，可加大此类项目的开发和推广力度；针对年轻游客居多的地区，打造更具时尚感和创新性的旅游产品，满足不同地域游客的多样化需求，提升 X 市旅游产品的竞争力。