# 上海商学院数字旅游实验室 第3期暑期训练营

**实验课一 旅游大数据用户画像分析**

# 实验课目标

本次实验课将会基于旅游消费数据对游客的消费进行画像分析，重点学习理解：

（1）学习如何对数据进行基本情况剖析及数据字段类型分析；

（2）学习如何对数据进行基本的分组统计分析，例如分析用户不同城市的旅游人群性别比例、不同城市的游客不同类型的人均消费对比等；

（3）学习如何对数据进行分布分析，例如分析某城市男性和女性在美食消费类别上的平均消费费用分布情况、不同城市旅游的人群对酒店住宿的满意度分布情况等。

# 数据介绍

本次实验课课程数据文件为“课程数据\_消费数据.xlsx”，总数据量为8000条。

本次实验课课程数据中的姓名、身份证号等字段均为虚拟数据，不涉及个人隐私信息。

# 课程一：数据字段基本情况分析

## 学习要点1：字符型字段枚举值数量分析

Excel函数学习：

从一个范围或数组返回唯一值：UNIQUE函数，语法如下：

UNIQUE(array,by\_col,exactly\_once)

▪ Array: 从中返回唯一行或列的范围或数组；

▪ By\_col: 是一个逻辑值: 比较各行并返回唯一值 = FALSE，或已省略；比较各列并返回唯一列 = TRUE；

▪ Exactly\_once: 是一个逻辑值: 从数组中返回只出现一次的行或列 = TRUE；从数组中返回所有不同的行或列 = FALSE，或已省略。

该函数用来判断一个字段的枚举值数量。

## 学习要点2：字符型字段枚举值数据量分析

Excel函数学习：

计算某个区域中满足给定条件的单元格数目：COUNTIF函数，语法如下：

COUNTIF(range,criteria)

▪ Range: 要计算其中非空单元格数目的区域；

▪ Criteria: 以数字、表达式或文本形式定义的条件。

该函数用来判断一个字段某个枚举值的数据数量。

## 学习要点3：数值型字段最大最小值计算

Excel函数学习：

计算数值型字段的最大值、最小值函数：MIN、MAX，语法如下：

MIN(number1,number2,...)

▪ Number1: number1,number2,... 是准备从中求取最小值的 1 到 255 个数值、空单元格、逻辑值或文本数值；

▪ Number2: number1,number2,... 是准备从中求取最小值的 1 到 255 个数值、空单元格、逻辑值或文本数值。

MAX(number1,number2,...)

▪ Number1: number1,number2,... 是准备从中求取最大值的 1 到 255 个数值、空单元格、逻辑值或文本数值

▪ Number2: number1,number2,... 是准备从中求取最大值的 1 到 255 个数值、空单元格、逻辑值或文本数值

MIN和MAX函数用来计算数值型字段的最大最小值。

## 课程作业1-1

**【作业1】**

根据实验课程演示，对于原始数据进行数据字段基本情况分析，并完成下表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **数据字段类型分析** | | | | | | | |
| **序号** | **字段名称** | **字段类型** | **字符型字段分析** | | | **数值型字段** | |
| **枚举值是否可穷尽** | **枚举值** | **枚举值数据量** | **最小值** | **最大值** |
| 1 | 姓名 | 字符型 | 否 | - | - | - | - |
| 2 | 身份证号 | 字符型 | 否 | - | - | - | - |
| 3 | 性别 | 字符型 | 是 | 女 | 4048 | - | - |
| 男 | 3952 | - | - |
| 4 | 民族 | 字符型 | 是 | 朝鲜族 | 1900 | - | - |
| 汉族 | 1924 | - | - |
| 回族 | 2092 | - | - |
| 蒙古族 | 2084 | - | - |
| 5 | 出生年月 | 日期型 | 否 | - | - | - | - |
| 6 | 旅游城市 | 字符型 | 是 | 杭州 | 1988 | - | - |
| 重庆 | 1964 | - | - |
| 深圳 | 1924 | - | - |
| 上海 | 2124 | - | - |
| 7 | 消费类型 | 字符型 | 是 | 美食 | 2000 | - | - |
| 纪念品 | 2000 | - | - |
| 酒店住宿 | 2000 | - | - |
| 交通出行 | 2000 | - | - |
| 8 | 人均消费额 | 数值型 | 否 | - | - | 10.43 | 499.44 |
| 9 | 满意度评价 | 数值型 | 否 | - | - | 4 | 10 |
| 10 | 天数 | 数值型 | 否 | - | - | 2 | 5 |

# 课程二：用户消费基本情况统计

## 学习要点1：分组进行字符型字段的计数统计

分析问题1：不同城市的旅游人群性别比例是多少？绘制饼图进行分析。

（1）要点

这里可以看到一共有4个城市：上海、深圳、杭州、重庆，两种性别比例：男、女。所以这个问题本质是以1个字符型字段进行分组，计数统计另一个字符型字段的计数，在这里则是以“城市”进行分组分析，每个组分别计算出男性游客和女性游客的人数。

（2）函数学习 – COUNTIFS

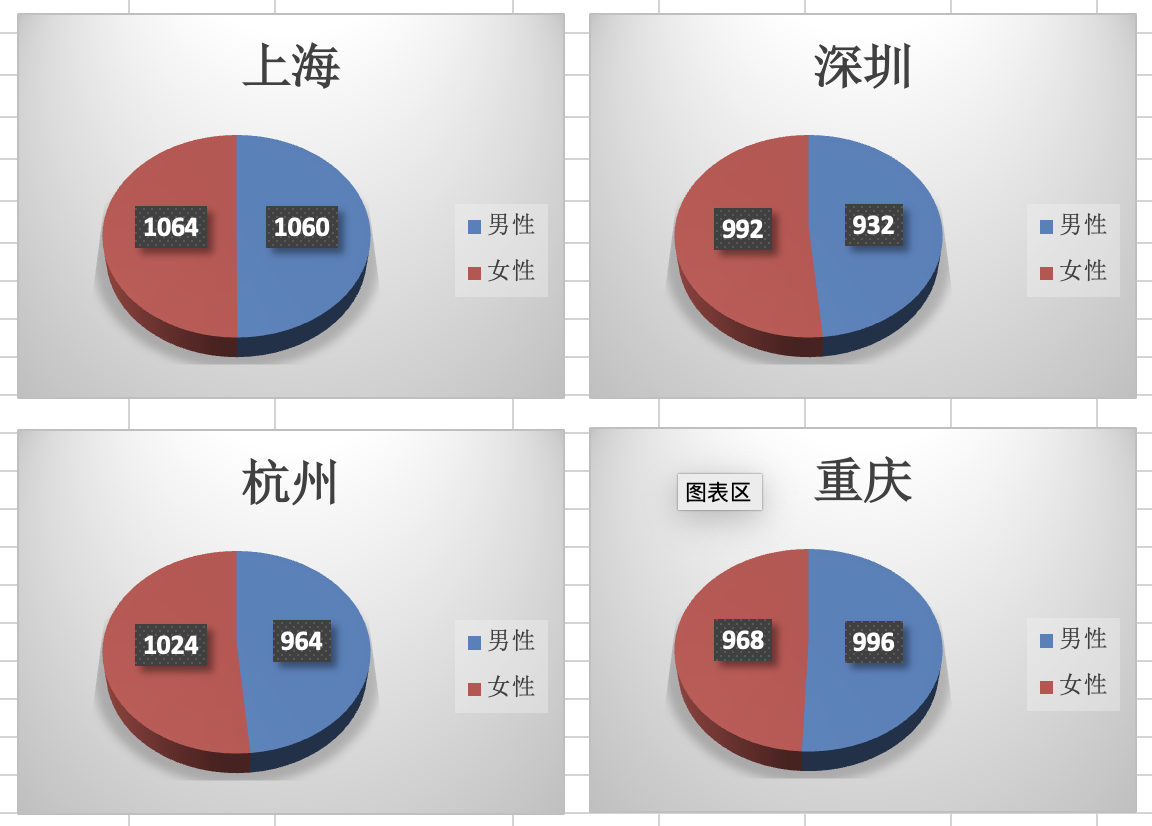
统计一组给定条件所指定的单元格数，语法如下：

COUNTIFS(criteria\_range,criteria,...)

▪ Criteria\_range1: 是要针对特定条件求值的单元格区域

（3）分析并完成以下图表

|  |  |  |
| --- | --- | --- |
| **不同城市的旅游人群性别比较** | | |
| **城市** | **男性** | **女性** |
| 上海 | 1060 | 1064 |
| 深圳 | 932 | 992 |
| 杭州 | 964 | 1024 |
| 重庆 | 996 | 968 |



## 学习要点2：分组进行数值型字段的统计

分析问题2：不同城市的游客不同类型的人均消费比较？绘制多系列柱状图进行分析。

（1）要点

这里可以看到一共涉及了3个字段：

* 字符型字段：4个城市：上海、深圳、杭州、重庆；
* 字符型字段：4种消费类型：美食、纪念品、酒店住宿、交通出行；
* 数值型字段：人均消费。

分组以“城市”字段进行分组，再对消费类型进行均值统计,注意了，所有的分组统计，都是基于一个字符型字段进行分组。

（2）函数学习 – AVERAGEIFS

查找一组给定条件指定的单元格的平均值(算术平均值)，语法如下：

AVERAGEIFS(average\_range,criteria\_range,criteria,...)

▪ Average\_range: 是用于查找平均值的实际单元格

▪ Criteria\_range1: 是要针对特定条件求值的单元格区域

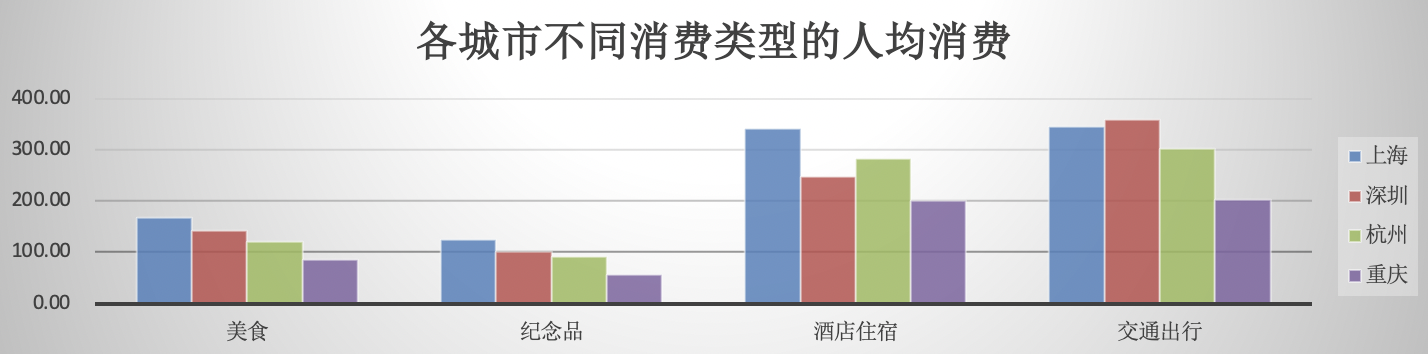
▪ Criteria1: 是数字、表达式或文本形式的条件，它定义了用于查找平均值的单元格范围

▪ Criteria\_range2: 是要针对特定条件求值的单元格区域

▪ Criteria2: 是数字、表达式或文本形式的条件，它定义了用于查找平均值的单元格范围

（3）分析并完成以下图表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **不同城市的游客不同类型的人均消费比较** | | | | |
| **城市** | **美食** | **纪念品** | **酒店住宿** | **交通出行** |
| 上海 | 167.50 | 125.17 | 340.27 | 343.93 |
| 深圳 | 141.14 | 100.34 | 246.69 | 359.17 |
| 杭州 | 119.61 | 91.22 | 283.00 | 302.54 |
| 重庆 | 84.58 | 56.64 | 200.40 | 203.26 |



## 课程作业1-2

**【作业1】**

基于4.1学习要点1，请分析不同城市（4个城市）不同消费类型（4种消费类型）的数据情况比较，并分别按城市绘制4个饼图进行分析。

**【作业2】**

基于4.2学习要点2，请分析不同城市（4个城市）不同消费类型（4种消费类型）的满意度情况比较，并绘制多系列柱状图。

# 课程三：用户消费满意度分析

在之前的两节课中，对于数据分析只是在统计层面（计数、求均值等），本节课要学习如果对数据进行分布分析。

分布分析是统计学中的一种工具，它用于描述和概括数据集中变量的分布特征。具体而言，分布分析可以帮助我们理解数据在不同数值范围内的分布情况，比如它可以表明一个变量的取值是否集中在某个特定区域，或者是否均匀分布在整个取值范围内。

分布分析常常涉及以下几个方面：

* 频数分布（Frequency Distribution）：通过统计每个取值或取值区间的频数（即出现次数）来描述数据的分布。
* 累计频数（Cumulative Frequency）：统计小于等于某个取值的所有频数的累计和。
* 频率分布（Frequency Distribution）：是频数分布的归一化形式，即每个取值或取值区间的频数占总数的比例。
* 累计频率（Cumulative Frequency）：频率的累计，它显示的是小于等于某一取值的所有频率的总和。
* 概率分布（Probability Distribution）：当数据集代表一个随机变量的时候，分布分析就涉及到概率分布，比如二项分布、正态分布、泊松分布等。

本次课程主要以**“频数分布”、“累计频数”**为学习内容，对于分布数据的可视化，大部分可以使用**直方图（Histograms）**或**箱形图（Box Plots）**的方式。

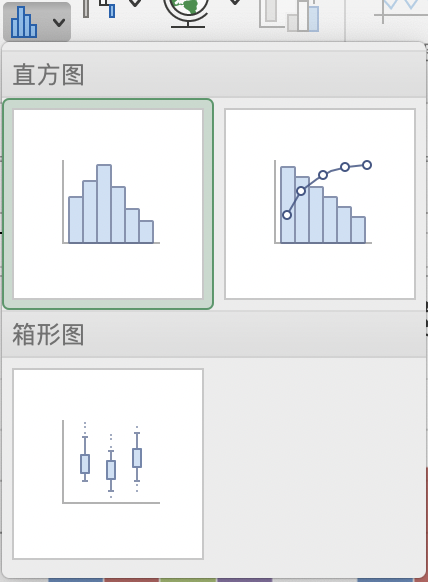
## 学习要点1：直方图绘制及分析

分析问题1：满意度评价整体数据分布情况如何？请用直方图进行分析。

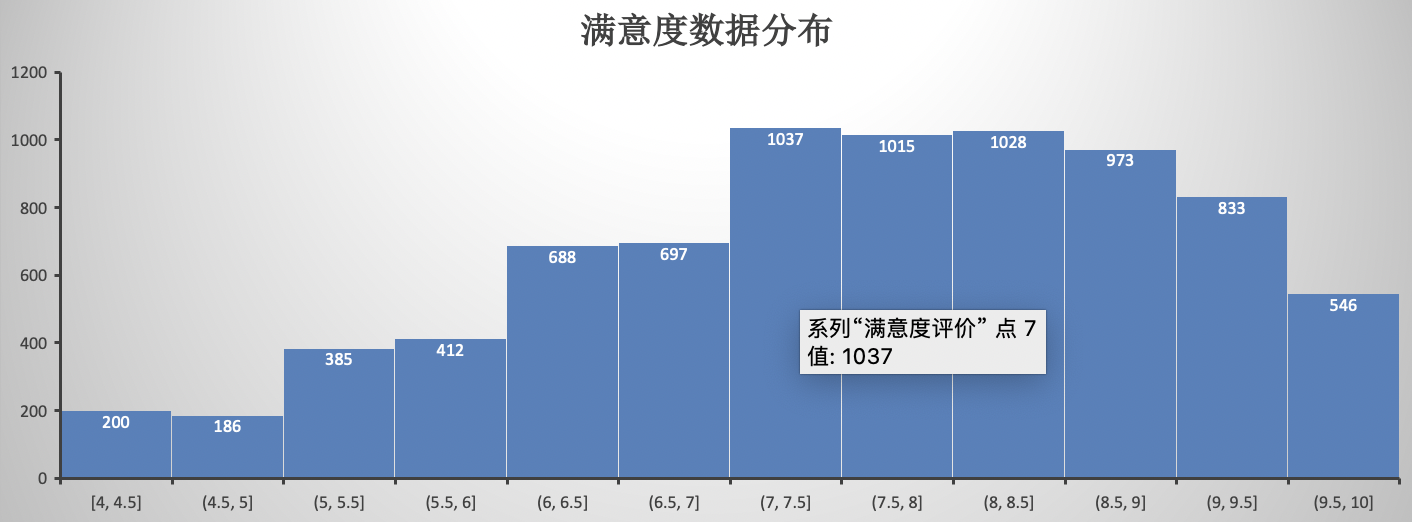
（1）要点

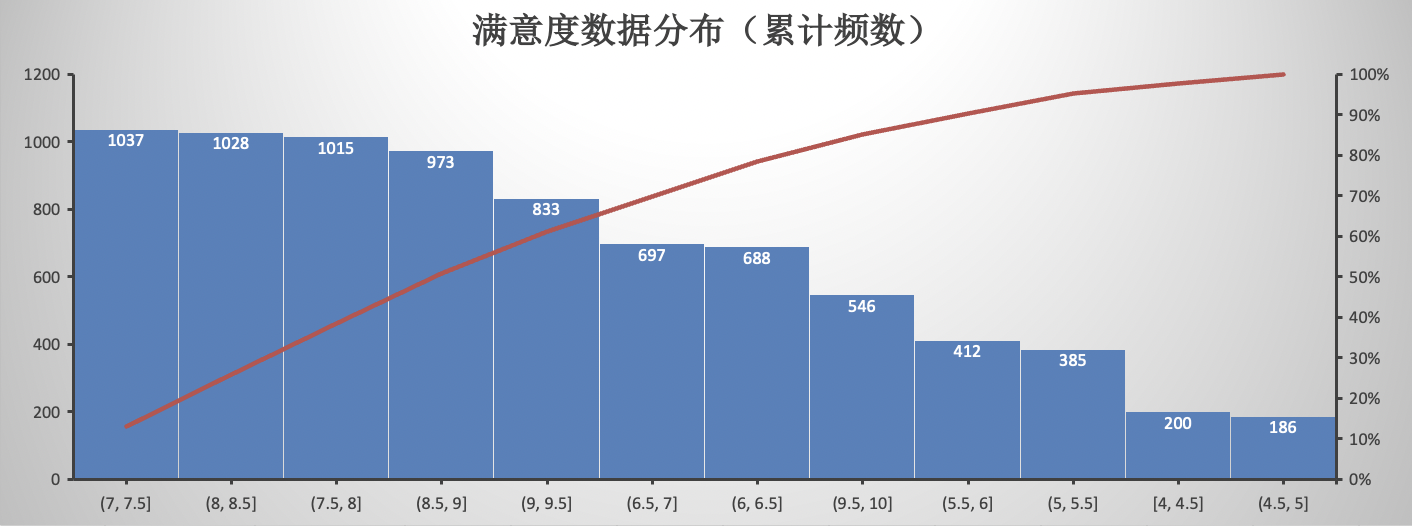
通过excel的直方图进行分析，学会通过对直方图进行箱属性的设置来查看结果。

（2）图表学习 – 直方图



（3）绘制直方图并进行分析





箱属性设置：按照0.5设置箱宽度。

按图分析：

* 整体来看，7-9分的评价是分布较多的满意度评分区域；

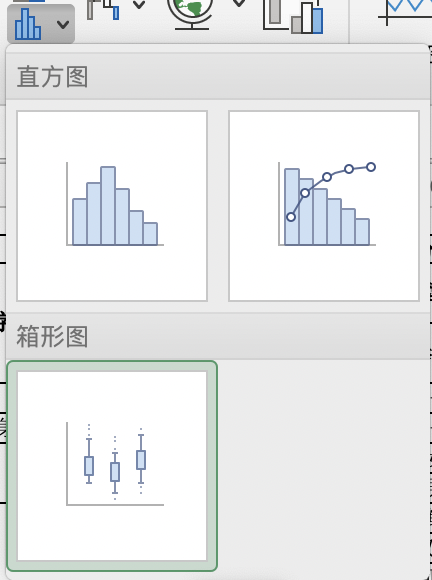
## 学习要点2：箱形图绘制及分析

分析问题2：按照不同城市分析满意度的分布情况如何？请用箱形图进行分析。

（1）要点

直方图对于单组数据进行分析较为直观，但这里“不同消费类型”或“不同城市”属于按照某个字符字段进行分组统计，这里就需要用到箱形图。

（2）图表学习 – 箱形图



（3）绘制箱形图并进行分析

图表, 箱线图

描述已自动生成

按图分析：

* 重庆整体满意度最高、杭州最低；
* 重庆满意度分布数据较集中，代表大家看法趋近一致，而杭州、深圳满意度分布较分散，代表大家不同观点较多；
* 重庆满意度的极小值比杭州满意度的均值高；
* 上海和深圳的满意度整体评价较为趋近，但上海整体略高一些。

## 课程作业1-3

**【作业1】**

基于5.1学习要点1，请分析人均消费的整体数据分布情况如何？请用直方图进行分析。

**【作业2】**

基于5.2学习要点2，按照不同消费类型或者城市，人均消费的的整体数据分布情况如何？请用箱形图进行分析。