演讲稿

1.

My name is Sarah, today I’m going talk about my Kaggle topic which is Hotel Tulip. It is a hypothetical organisation.

2.

I will briefly introduce this hotel’s background. Additionally, I'll break down my analysis of this hypothetical hotel dataset into two parts.

3.4

First of all, this hotel is a hypothetical organisation, like I just mentioned. It aims to train hoteliers, a hotel with creativity and the spirit of a TULIP lab. This hotel stores all and if the hotel website's web traffic and data about the various relevant web pages used by users.

5.

The aim is to discover potential users, improve the online experience of potential users and help the marketing department to identify and uncover customer behaviour patterns. The dataset used to discover patterns of user visits to different web pages is also in web log format.

首先，这个酒店是一个假想组织，像我刚刚提到的。它的目的在于培养酒店经营者，不仅如此它也是一个具有创造力和TULIP实验室精神的酒店。这个酒店将所有登陆过酒店网站的网络流量和用户使用的各种相关网页有关数据都进行了储存。

目的是为了发现潜在用户，改善潜在用户的在线体验，帮助市场部门识别和挖掘客户的行为模式。为了发现不同网页网页用户访问的模式，我们用到的数据集也是web日志格式的。

6.7

1.将得到的数据集进行提取：获取到的数据集是压缩包的形式，先要进行解压。得到了图一所展现的所有压缩包里日志格式的数据集名称，并进行统计，知道一共有120个文件。

8

在读取所有的文件后，将数据进行合并并打印前五行进行一般检查，这里的目的就是看看表头和数据有没有一一对应上。而且在查看的同时可以表头中的信息。

9.

这里我做了一个数据信息属性的表，可以给大家解释一下，做数据分析时主要用到的几个属性。日期时间不用说，sc-status的意思是协议状态。简单来讲就是访问网站是否成功，或者没有成功。C-ip是用户访问的ip地址，从这里可以分析道访问网站的用户是从那里访问的，可以从国家或者城市划分。

10.

这时候数据还不是可用的，因为还没有被清理。通过计算，将每列中数据带有nan或者空白超过百分之15的数据进行整列删除。同时，我们可以看到哪些没有超过百分之15的数据中其实也有空白或者nan这个时候就单个删除行。最后清理完成的数据是8百万434645行

11.12

进行分析：首先我对点击流量进行了分析。这里主要针对每个小时来进行统计。讲每个小时相同的行算作流量绘制出来的图。可以得到一些非常有趣的信息，由于时差关系这里采用的是澳大利亚墨尔本时间，可以看到再墨尔本时间的超过40万以上的访问量再凌晨的2-9点，下午1-3又迎来了一个小高峰。同时再晚上7点到11点这个时间段访问量较少

13.

第二步我对用户访问页面的协议状态进行了统计。访问成功的人占据大多数达到了600万.而呈现304状态排第二，304的意思就是说服务器对网页有缓存。那么证明这些用户是第二次或者多次访问过该网站。

14.

第三步我对用户登陆网站的ip进行分析，主要从国家和城市分析。想要了解潜在用户所在地的可能性。首先是国家。我发现香港，中国和美国是访问前三的国家。香港占据大多数，那么再市场营销方面可以主要针对这三个国家。

15.

其次再看看城市，发现这三个国家的一线城市是主要访问者比入中国的北京和美国的纽约。

16.17

简单回顾一下我做这四部分析的目的在于想要了解用户在那个时间段访问的频率比较高，他们访问的频率是怎样的，是第一次访问还是回头客。从地理上分析的主要目的是发现潜在用户所在地理位置，开拓市场，并及时调整组织发展战略和规划。