GROUP PROJECT FINAL REPORT

Research On Doctor-Patient Behavior And Medical Resource Allocation Of Internet Medical Platform

廖家瑞 (Jiarui Liao) 1801212884
刘德明 (Deming Liu) 1801212889
曹阳 (Yang Cao) 1801212825
王如思 (Rusi Wang) 1801212938
常鑫磊 (Xinlei Chang) 1801212779

1 Project Introduction

In recent years, China's medical and health undertakings have developed significantly, but this development also faces some challenges. For example, there is information asymmetry between doctors and patients, and the uneven distribution of medical resources between regions and between urban and rural areas. These phenomena have led to contradictions between doctors and patients and problems such as "inadequate and overly expensive medical service" in large hospitals. Specifically, compared with developed countries, China has been suffering from serious shortage of medical resources and unreasonable allocation of medical resources for a long time. According to the statistics of our ministry of health, our country has a huge population, accounting for more than one fifth of the world's population, but we only have two percent of the world's medical resources.

Internet medical treatment was created in order to solve these problems under the technical background of the time. Internet medical treatment will help patients search for information, and it will also facilitate the reorganization and sharing of resources such as doctors' knowledge and skills in space and time. Therefore, Internet medical treatment will greatly change the traditional medical model, and have a certain impact on patients' medical decision. How to use information and network technology to solve the imbalance of supply and demand and distribution of medical resources has become a hot issue in the field of research and practice.

This project will be carried out with multi-dimensional Internet data of the Haodf medical platform. **Established in 2006, Haodf medical platform is the leading Internet medical platform in China.** (https://www.haodf.com) We will crawl the data from 2008 to 2018 on Haodf website, and then carry out subsequent cleaning, analysis and visualization. The data set will contain all hospitals of Shanghai city department of the relevant data, including basic information on doctors and patients, the geographical position, participation, including the time of registration, login, and stay logged in time, number of posts, etc.), post (post, reply, questions), the content of the post (the length of the posts, posts, the theme of the emotion), evaluation data, etc. This project will examine the impact of geographic factors and willingness to consult from the perspective of doctor-patient interaction (reflected in doctor-patient counseling and response). The project will focus on the improvement effect of the Internet on the imbalance of medical resources, and also plan to visualize the geographical distribution of medical allocation on the Internet medical platform. Finally, this project hopes to put forward corresponding suggestions on the influence of Internet medical treatment on doctor-patient behavior and resource allocation.

2 Data Description

2.1 Data overview

On the website of Haodf, we collected all relevant data from most departments of all hospitals in some regions (Shanghai, Anhui and Gansu were sampled accordingly) The data includes basic information on doctors and patients, the geographical position, participation, (including the time of registration, login time, stay time, number of posts, etc), post (post, reply, ask questions), the content of the post (the theme of the posts, the length of the posts, the emotion), evaluation data, etc.

Generally speaking, the data we collected in the project can be divided into three parts.

The first part includes hospital information, such as location, grade, department name, etc (as shown in **Figure 1.**).

The second part includes the information of doctors, such as location, title and service status (as shown in **Figure 2.**).

The third part includes patient information, such as: location, consultation times, consultation length of time, etc (as shown in **Figure 3.**).

2.2 Intrinsic big data properties

The data of this project conforms to the inherent characteristics of big data.

1) **Variety:** both structured and unstructured. Structured data, such as the age of the patient, the title of the doctor, the number of consultants, etc. Unstructured data mainly includes the content of consulting details.

2) **Volume:** the project has collected relevant data of all Shanghai hospitals on the website of Haodf for nearly ten years, with the total number of tens of millions, and the storage capacity is TB, which is consistent with the characteristics of large data Volume.

3) **Velocity:** project data includes doctor-patient communication data all the time. The data changes quickly and must be processed quickly.

4) **Value:** the data of the project itself has a low Value density, but if appropriate processing is carried out, valuable information can be mined, such as the improvement effect of Internet medical model on the allocation of medical resources.

2.3 Relational databases

This project plans to use a **relational database: Microsoft Access** to store data. The Access database can connect the data tables used by the project to obtain the table of research objects, representing the relationship between the data. The connection relationship between the tables is shown in **Figure 4**.

According to the data we collected in 2.1, we split the data into six tables: *Hospitals, Departments, Doctors, Consults, Patients, Disease_for_department* (as shown in **Figure 4** in the appendix). The database is designed based on 2NF (considering the difficulty of SQL query, we did not implement 3NF, because there might be more than 1 primary key in a table, which may bring difficulty when designing SQL query statements.)

2.4 Data preprocessing

There is a lot of messy code and data redundancy in the data, so it is necessary to preprocess the data before the formal analysis (as shown in **Figure 5** and **Figure 6**).

3 Data Visualization

3.1 Main Idea

After data processing, we first perform data visualization. Good geographic information display can directly highlight the problem and also help to better predict the possible results. We first make a heat map of consulting Shanghai doctors across the country. And by collecting GDP per capita, the number of patients, and the number of netizens to intuitively test our hypothesis that the number of consultants is related to GDP and the number of Internet users. Due to the large amount of data, it is difficult to analyze the national data in a short time, so we chose to extract relevant data from Anhui Province and Shanghai for detailed analysis. Based on the results of data visualization, then we will mainly analyze the impact of Shanghai's medical level and other factors on Anhui patients' choice of local or Shanghai medical treatment. In the process of data visualization, we used the Power Map plug-in of Excel, the business intelligence software Tableau, and the Python drawing package Pyecharts.

3.2 Nationwide Analysis

Cluster chart (as shown in **Figure 7**): We combine the per capita GDP, the number of patients, and the number of netizens, and then normalize them in this Cluster diagram. It’s to intuitively test our hypothesis that the number of consultants is related to GDP and the number of Internet users.

Heat map (as shown in **Figure 8**): This map can intuitively reflect the quantity and intensity of distribution through color and geographic location. It can be clearly seen that the distribution of consulting patients is closely related to distance and population density.

Combining the two graphs above, we can get several conclusions that: (1) The closer the area is to Shanghai, the greater the number of people seeking medical treatment in Shanghai. Obviously, Shanghai has the largest number. (2) The number of people consult Shanghai doctors is also related to GDP per capita. For example, Guangdong has a higher GDP per capita. Although it is far away from Shanghai, there are more people than other regions. (3) The choice of patients is also related to the Penetration of the network. The higher the proportion of Internet users, the corresponding number of people who use Haodf online is larger compared with the regions where distance and per capita GDP are close.

3.3 Province Level Analysis

We draw two bar charts on the doctor and patient level statistics of Anhui Province in consult list, and four regional maps of the region feature of Anhui Province.

The bar charts (as shown in **Figure 9** and **Figure 10**) show that:

(1) Although the departments are different, the main part of the consultations is to Grade A hospital.

(2) From the perspective of departments, the number of pediatric consultations occupies a

major advantage. Dermatology, surgery and internal medicine are for the main proportion of consultation. This conclusion is related to the characteristics of the disease.

(3) Patients prefer to consult Chief, Associate Chief, Attending Doctors online.

The regional maps (as shown in **Figure 11-14**) give an intuitive observation of the features of cities in Anhui Province. Through the analysis of the medical level scores, GDP per capita of cities in Anhui and distance and difference of medical level between Shanghai and cities in Anhui. Then we can have a unified understanding of the medical level and GDP of each city in Anhui Province. This information could help us in the model construction.

4 Model and Implementation

4.1 PCA

Principal component analysis is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables.

Since our purpose is to find the factors that influence the patient behavior, we use principal components analysis to construct the medical level indicator. Based on many medical references and data we collected, we decide constructing the medical level indicator by using eight features which include the elderly dependency ratio, Number of medical institutions, Number of hospital, Number of medical beds, Number of medical personnel, Number of practicing physician, Average hospitalization day and GDP.

The scree plot shows that three components can explain the 81.5% of total variance, which is enough for our analysis. We also find the intuitive meaning of these three components, we can see that the first components is highly correlated with GDP and medical facilities indicator, so the first component reflects the economic and medical resource level. The second components is highly corelated with the number of hospital and the number of medical institutions, which reflect the degree of medical competition. The third component is highly correlated with the elderly dependency ratio, which reflect the population and aging factor.

After contracting all the indicator to the three components, we construct the Medical level using these three components, the weight of each components can be calculated by the following formula:

$$Weight_i = \frac{Var(x_i)}{\sum_{i=1}^{3} Var(x_i)} = \frac{\lambda_i}{\sum_{i=1}^{3} \lambda_i}$$

$$Medical\ level = \sum_{i=1}^{3} Weight_i * Component_i$$

4.2 Logistic Regression

After using PCA to form 3 principle components, we add the weighted value of each component to form a new variable called Medical_Level. It means the comprehensive score of the city's medical level. And we combine Medical_Level and distance to Shanghai and GDP of this city into the patient's data. However, every patient now has the same city attributes, it is hard for Logistic Regression Model to discriminate the samples.

Therefore, we investigate the other information and invent a new variable called Utility Score. It measures the urgency or utility of a patient. For example, if a patient uses more words to describe his illness, the Utility Score will be higher. Also, if a patient makes a phone call to the doctor, the Utility Score will be higher too. In general, we define Utility Score as a comprehensive score of Length of Description, Numbers of Phone Calls, Total Time of Consulting.

Then we use Logistic Regression to process our data. The independent variables are city's distance to Shanghai, GDP, Medical_Level and the Utility Score of each patient. The dependent variable is whether the patient chooses Shanghai's hospitals. If it is true, it will be 1, otherwise 0. The distribution formula as follows:

$$P_{Y=1} = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}}$$

The result shows that the accuracy, precision and recall are all about 60%, and the coefficients lead to our final results.

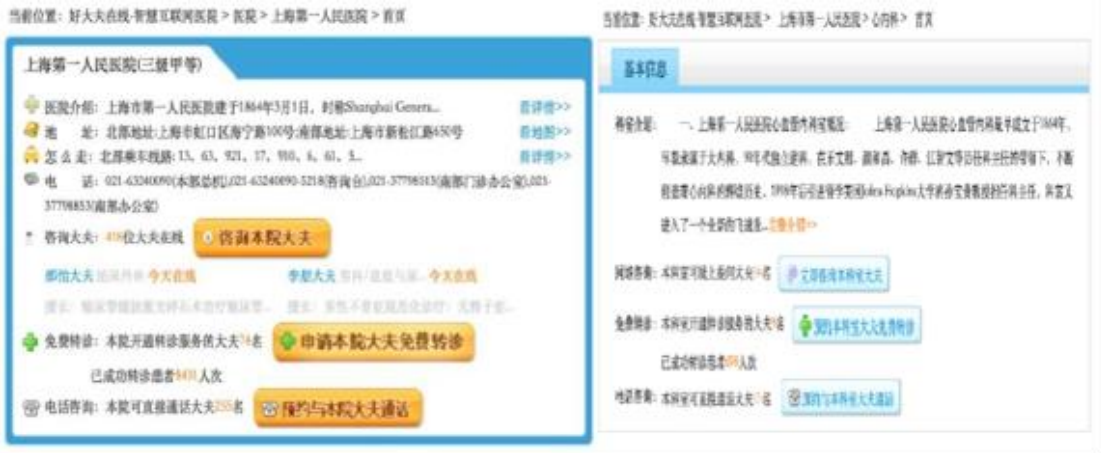
5 Summary and Outlook

The final question of our project is whether Internet medical platform is conducive to the allocation of medical resources? Our answer is yes! Taking Shanghai as an example, it can be seen from the visualization results that the consulting patients are all over the country, and people all over the country can enjoy the high-quality medical resources of Shanghai (especially Grade-A tertiary hospital) to a certain extent.

Regional factors and willingness to consult affect patients' decision-making. (1) The closer the distance is, (2) the higher the GDP is, and (3) the poor the medical level is, the more patients choose Shanghai, and (4) the stronger their willingness to consult, the more they choose Shanghai. That may because of the convenience of distance, the more abundant funds and the stronger willingness of better hospitals and doctors.

What else can we do in the future? The data can be combined with the patient's personal information (family income, duration of illness, age) to provide a more detailed analysis of the factors influencing the patient's decision. What's more, time variables can be added into the model to study the changes of patients' decision-making in different periods.

Appendix



当前位置：好大夫在线-智慧互联网医院 > 医院 > 上海第一人民医院 > 首页

上海第一人民医院(三级甲等)

医院介绍：上海市第一人民医院建于1864年3月1日，时称Shanghai Genera... 查详情>>

地　　址：北部地址:上海市虹口区海宁路100号;南部地址:上海市新松江路650号 查地图>>

怎么走：北部乘车线路:13、63、921、17、910、6、61、5... 查详情>>

电　　话：021-63240090(本部总机);021-63240090-5218(咨询台);021-37798513(南部门诊办公室);021-37798853(南部办公室)

咨询大夫：418位大夫在线 咨询本院大夫

郭怡大夫 [illegible] 今天在线　　李毅大夫 [illegible] 今天在线

擅长：[illegible]　　擅长：[illegible]

免费转诊：本院开通转诊服务的大夫74名 申请本院大夫免费转诊

已成功转诊患者6431人次

电话咨询：本院可直接通话大夫255名 预约与本院大夫通话

当前位置：好大夫在线-智慧互联网医院 > 上海市第一人民医院 > 心内科 > 首页

基本信息

科室介绍：一、上海第一人民医院心血管内科室概况　上海第一人民医院心血管内科最早成立于1964年，[illegible] 1996年[illegible]Johns Hopkins大学[illegible]，科室又进入了一个全新的飞速发... 全部介绍>>

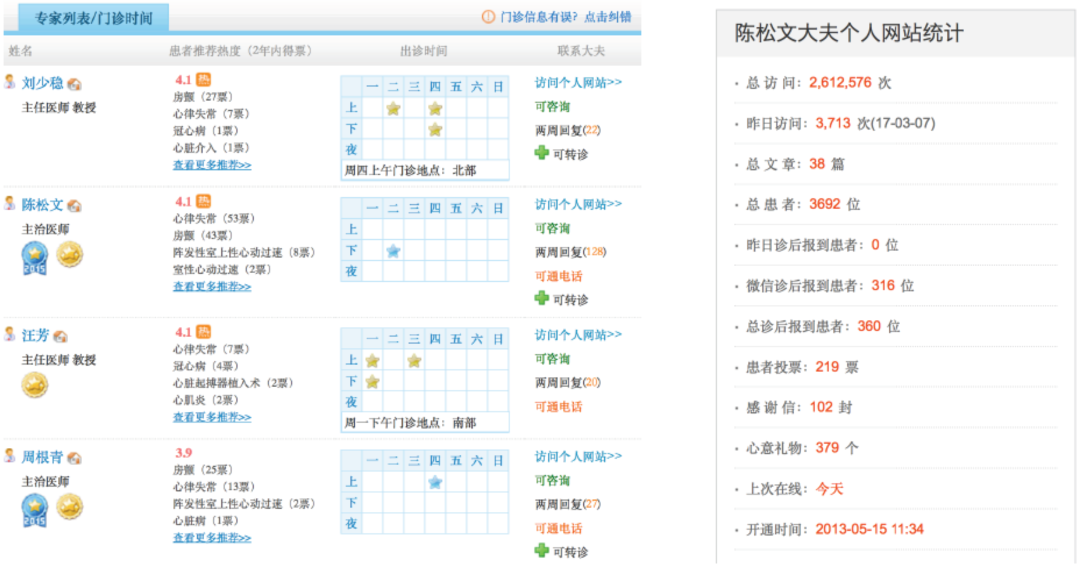
网络咨询：本科室可线上咨询大夫[illegible]名 在线咨询本科室大夫

免费转诊：本科室开通转诊服务的大夫[illegible]名 预约本科室大夫免费转诊

已成功转诊患者[illegible]人次

电话咨询：本科室可直接通话大夫[illegible]名 预约与本科室大夫通话

Figure 1: Hospital information interface (Source: Haodf website)



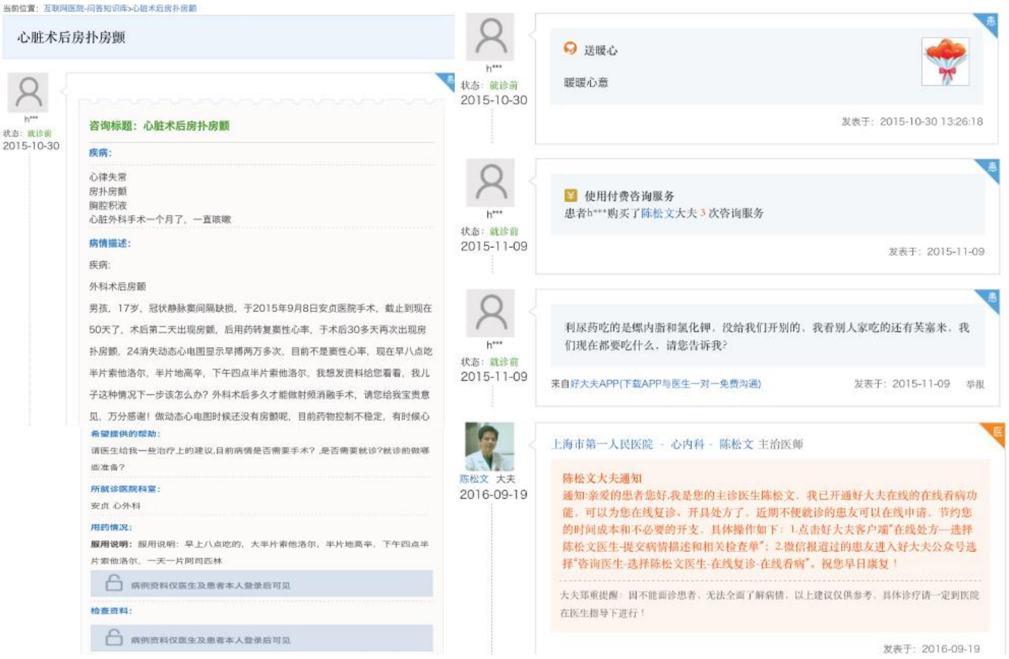
专家列表/门诊时间　　ⓘ 门诊信息有误？点击纠错

姓名	患者推荐热度（2年内得票）	出诊时间	联系大夫
刘少稳 主任医师 教授	4.1 热 房颤（27票） 心律失常（7票） 冠心病（1票） 心脏介入（1票） 查看更多推荐>>	上：二 ★，四 ★ 下：四 ★ 周四上午门诊地点：北部	访问个人网站>> 可咨询 两周回复(22) ✚ 可转诊
陈松文 主治医师	4.1 热 心律失常（53票） 房颤（43票） 阵发性室上性心动过速（8票） 室性心动过速（2票） 查看更多推荐>>	下：二 ★	访问个人网站>> 可咨询 两周回复(128) 可通电话 ✚ 可转诊
汪芳 主任医师 教授	4.1 热 心律失常（7票） 冠心病（4票） 心脏起搏器植入术（2票） 心肌炎（2票） 查看更多推荐>>	上：一 ★，三 ★ 下：一 ★ 周一下午门诊地点：南部	访问个人网站>> 可咨询 两周回复(20) 可通电话
周根青 主治医师	3.9 房颤（25票） 心律失常（13票） 阵发性室上性心动过速（2票） 心脏病（1票） 查看更多推荐>>	上：四 ★	访问个人网站>> 可咨询 两周回复(27) 可通电话 ✚ 可转诊

陈松文大夫个人网站统计

- 总 访 问：2,612,576 次
- 昨日访问：3,713 次(17-03-07)
- 总 文 章：38 篇
- 总 患 者：3692 位
- 昨日诊后报到患者：0 位
- 微信诊后报到患者：316 位
- 总诊后报到患者：360 位
- 患者投票：219 票
- 感 谢 信：102 封
- 心意礼物：379 个
- 上次在线：今天
- 开通时间：2013-05-15 11:34

Figure 2: Doctor information interface (Source: Haodf website)



当前位置：互联网医院-问答知识库>心脏术后房扑房颤

心脏术后房扑房颤

h***
状态：就诊前
2015-10-30

咨询标题：心脏术后房扑房颤

疾病：

心律失常
房扑房颤
胸腔积液
心脏外科手术一个月了，一直咳嗽

病情描述：

疾病：

外科术后房颤

男孩，17岁，冠状静脉窦间隔缺损，于2015年9月8日安贞医院手术，截止到现在50天了，术后第二天出现房颤，后用药转复窦性心率，于术后30多天再次出现房扑房颤，24消失动态心电图显示早搏两万多次，目前不是窦性心率，现在早八点吃半片索他洛尔，半片地高辛，下午四点半片索他洛尔，我想发资料给您看看，我儿子这种情况下一步该怎么办？外科术后多久才能做射频消融手术，请您给我宝贵意见，万分感谢！做动态心电图时候还没有房颤呢，目前药物控制不稳定，有时候心

希望提供的帮助：

请医生给我一些治疗上的建议,目前病情是否需要手术？,是否需要就诊?就诊前做哪些准备？

所就诊医院科室：

安贞 心外科

用药情况：

服用说明： 服用说明：早上八点吃的，大半片索他洛尔，半片地高辛，下午四点半片索他洛尔，一天一片阿司匹林

病例资料仅医生及患者本人登录后可见

检查资料：

病例资料仅医生及患者本人登录后可见

h***
状态：就诊前
2015-10-30

送暖心

暖暖心意

发表于：2015-10-30 13:26:18

h***
状态：就诊前
2015-11-09

使用付费咨询服务

患者h***购买了陈松文大夫 3 次咨询服务

发表于：2015-11-09

h***
状态：就诊前
2015-11-09

利尿药吃的是螺内脂和氯化钾，没给我们开别的，我看别人家吃的还有芙塞米，我们现在都要吃什么，请您告诉我?

来自好大夫APP(下载APP与医生一对一免费沟通)　发表于：2015-11-09　举报

陈松文　大夫
2016-09-19

上海市第一人民医院 - 心内科 - 陈松文 主治医师

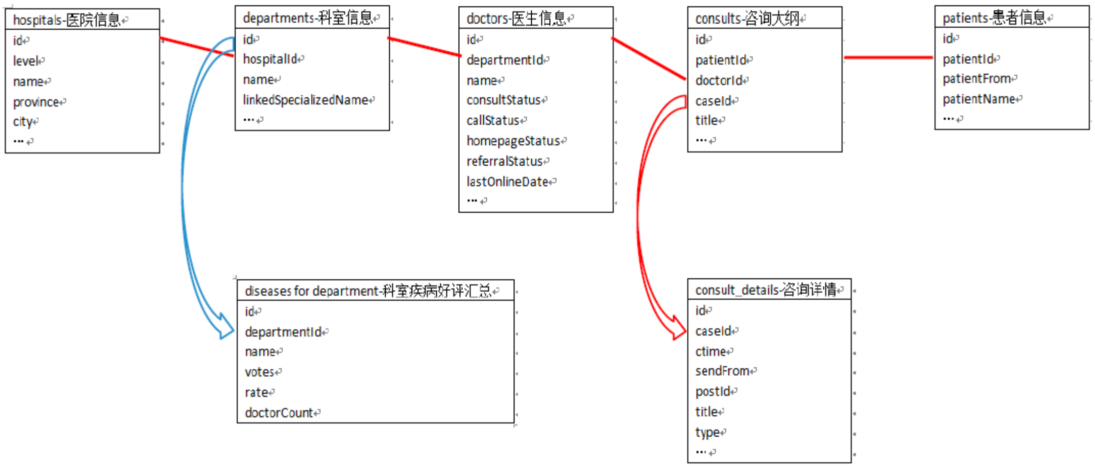
陈松文大夫通知

通知:亲爱的患者您好,我是您的主诊医生陈松文，我已开通好大夫在线的在线看病功能，可以为您在线复诊、开具处方了，近期不便就诊的患友可以在线申请，节约您的时间成本和不必要的开支，具体操作如下：1.点击好大夫客户端"在线处方—选择陈松文医生-提交病情描述和相关检查单"；2.微信报道过的患友进入好大夫公众号选择"咨询医生-选择陈松文医生-在线复诊-在线看病"，祝您早日康复！

大夫郑重提醒：因不能面诊患者，无法全面了解病情，以上建议仅供参考，具体诊疗请一定到医院在医生指导下进行！

发表于：2016-09-19

Figure 3: Patients information interface (Source: Haodf website)



hospitals-医院信息
id
level
name
province
city
…

departments-科室信息
id
hospitalId
name
linkedSpecializedName
…

doctors-医生信息
id
departmentId
name
consultStatus
callStatus
homepageStatus
referralStatus
lastOnlineDate
…

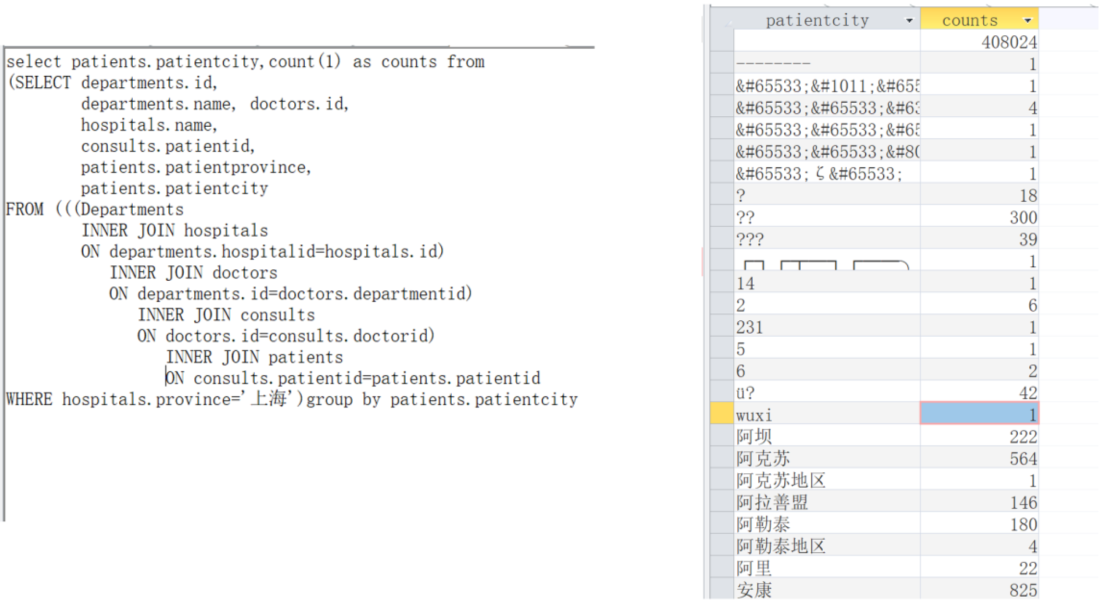
consults-咨询大纲
id
patientId
doctorId
caseId
title
…

patients-患者信息
id
patientId
patientFrom
patientName
…

diseases for department-科室疾病好评汇总
id
departmentId
name
votes
rate
doctorCount

consult_details-咨询详情
id
caseId
ctime
sendFrom
postId
title
type
…

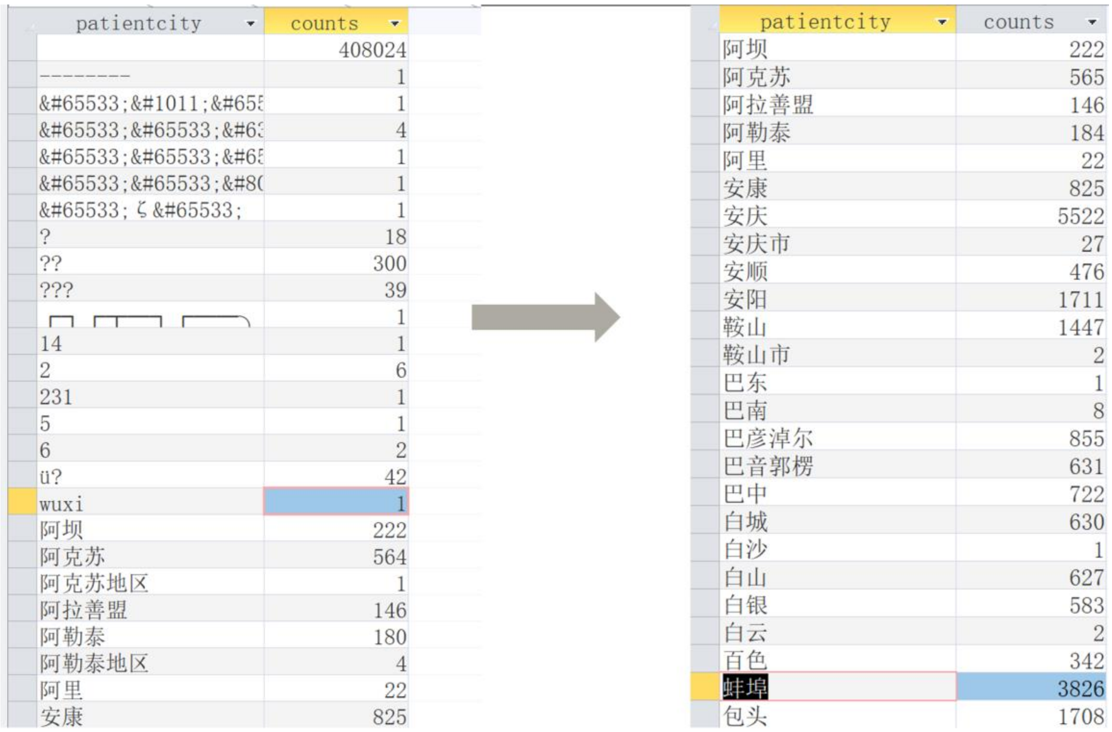
Figure 4: Relational Database based on 2NF



```
select patients.patientcity,count(1) as counts from
(SELECT departments.id,
        departments.name, doctors.id,
        hospitals.name,
        consults.patientid,
        patients.patientprovince,
        patients.patientcity
FROM (((Departments
        INNER JOIN hospitals
        ON departments.hospitalid=hospitals.id)
           INNER JOIN doctors
           ON departments.id=doctors.departmentid)
              INNER JOIN consults
              ON doctors.id=consults.doctorid)
                 INNER JOIN patients
                 ON consults.patientid=patients.patientid
WHERE hospitals.province=’上海’)group by patients.patientcity
```

patientcity	counts
	408024
--------	1
&#65533;&#1011;&#655	1
&#65533;&#65533;&#63	4
&#65533;&#65533;&#65	1
&#65533;&#65533;&#80	1
&#65533;ζ&#65533;	1
?	18
??	300
???	39
[illegible]	1
14	1
2	6
231	1
5	1
6	2
ü?	42
wuxi	1
阿坝	222
阿克苏	564
阿克苏地区	1
阿拉善盟	146
阿勒泰	180
阿勒泰地区	4
阿里	22
安康	825

Figure 5: Using SQL query code shown on the left, we can see there exists messy code and redundancy in the data.



patientcity	counts
	408024
--------	1
&#65533;&#1011;&#655	1
&#65533;&#65533;&#63	4
&#65533;&#65533;&#65	1
&#65533;&#65533;&#80	1
&#65533;ζ&#65533;	1
?	18
??	300
???	39
[illegible]	1
14	1
2	6
231	1
5	1
6	2
ü?	42
wuxi	1
阿坝	222
阿克苏	564
阿克苏地区	1
阿拉善盟	146
阿勒泰	180
阿勒泰地区	4
阿里	22
安康	825

patientcity	counts
阿坝	222
阿克苏	565
阿拉善盟	146
阿勒泰	184
阿里	22
安康	825
安庆	5522
安庆市	27
安顺	476
安阳	1711
鞍山	1447
鞍山市	2
巴东	1
巴南	8
巴彦淖尔	855
巴音郭楞	631
巴中	722
白城	630
白沙	1
白山	627
白银	583
白云	2
百色	342
蚌埠	3826
包头	1708

Figure 6: After data preprocessing, the data is cleaner and can be used in further analysis.

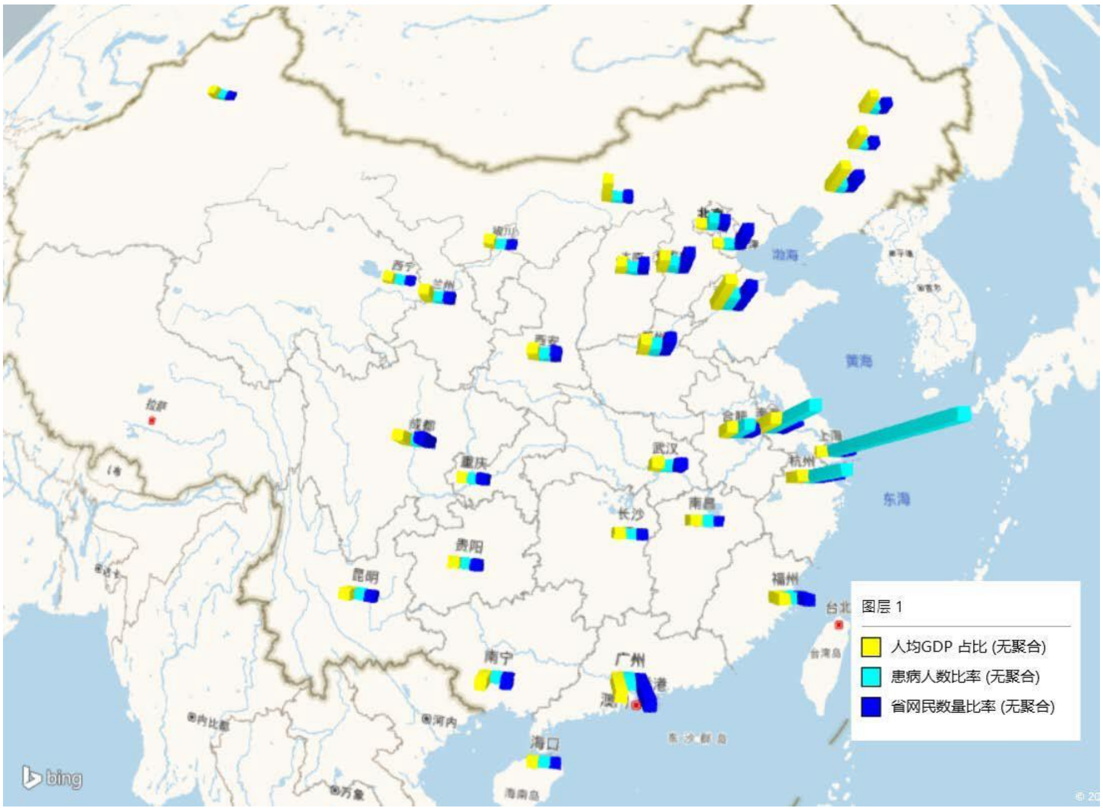


Figure 7: Cluster chart of province level features

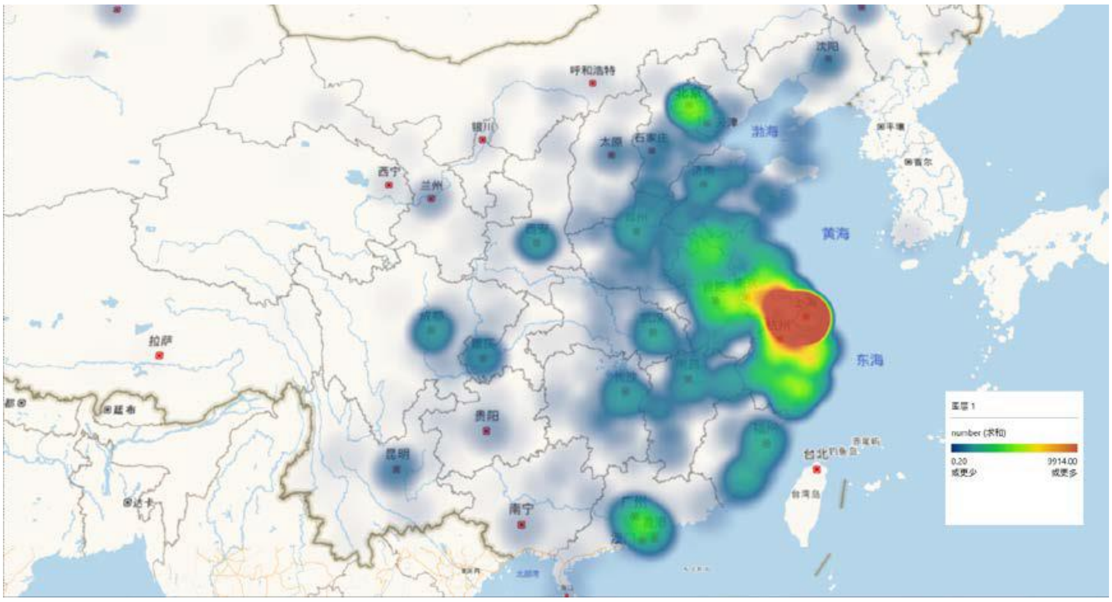


Figure 8: Heat Map of online consult to Shanghai

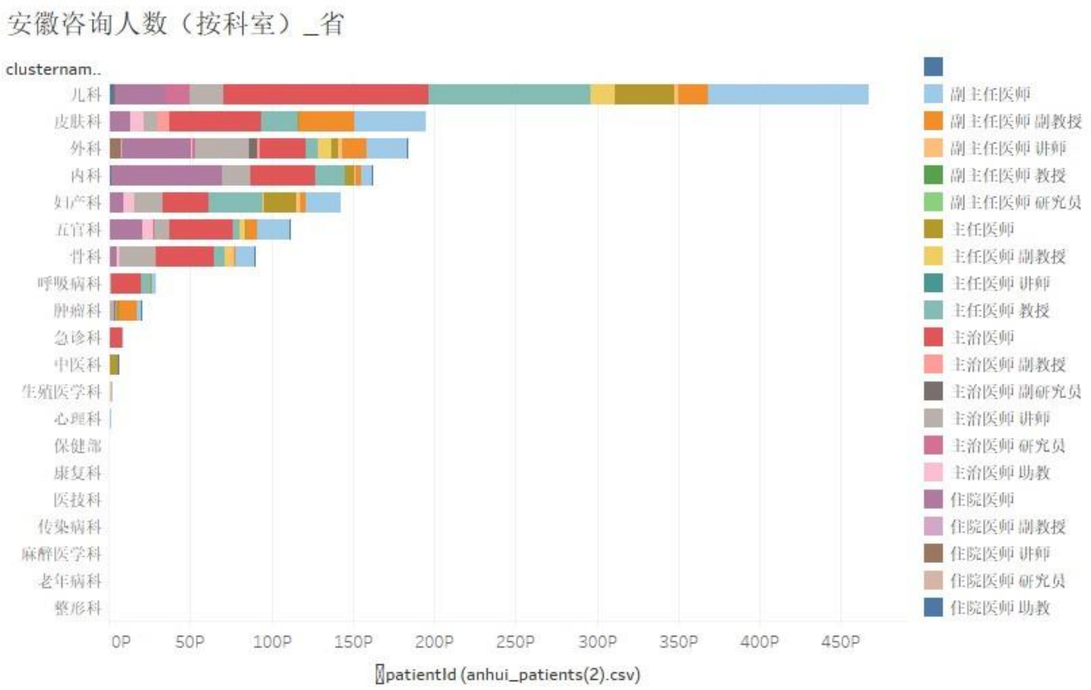


Figure 9: Doctor level statistics of Anhui in consult list

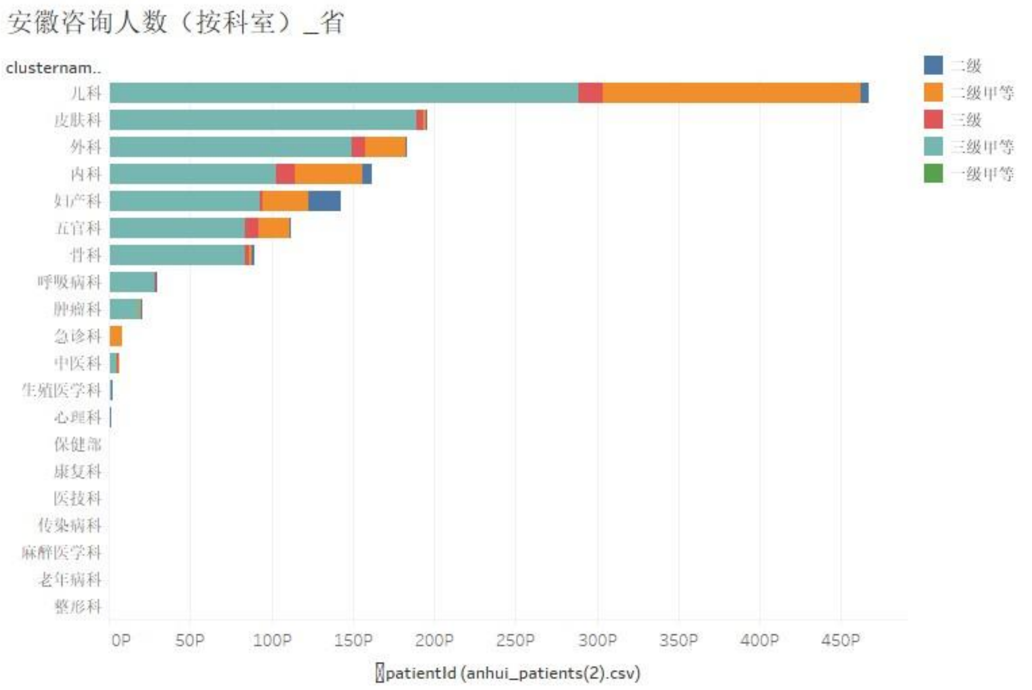


Figure 10: Hospital level statistics of Anhui in consult list

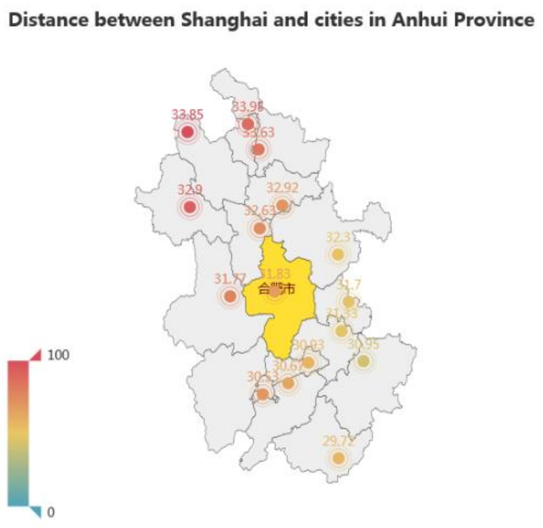


Figure 11: The region feature of Anhui Province (I)

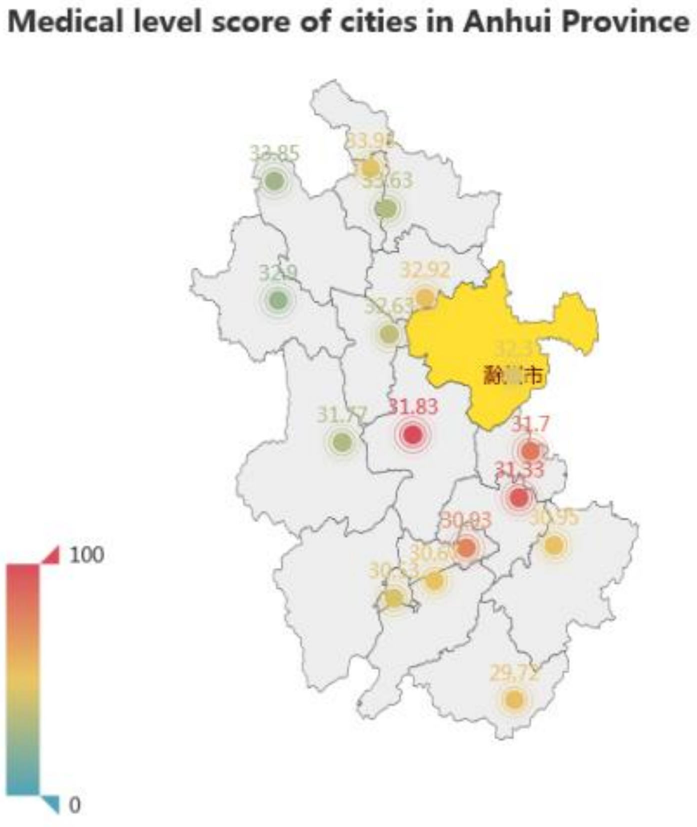


Figure 12: The region feature of Anhui Province (II)

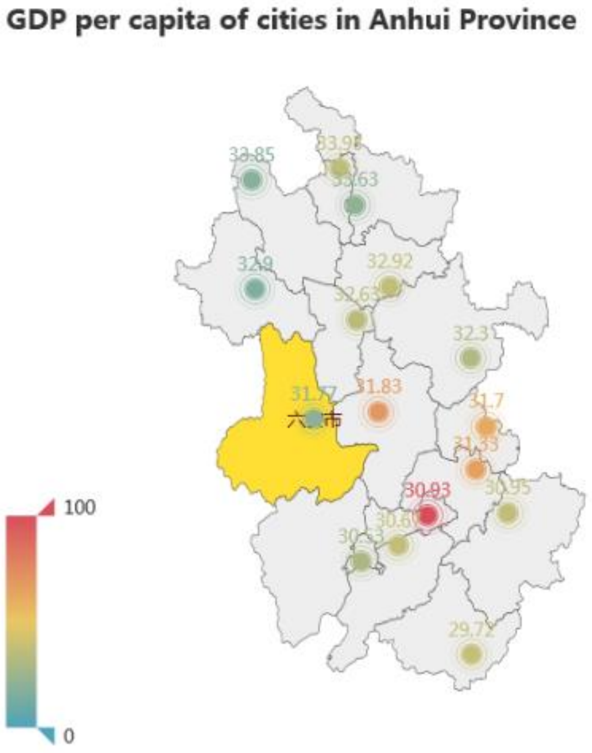


Figure 13: The region feature of Anhui Province (III)

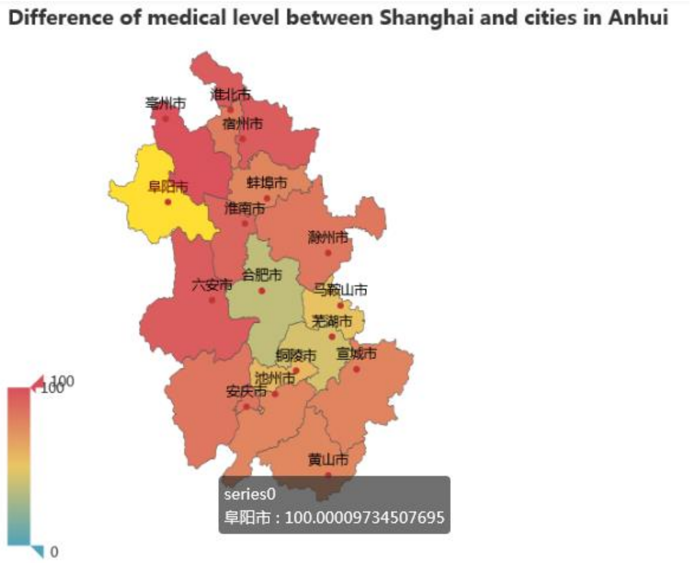


Figure 14: The region feature of Anhui Province (IV)

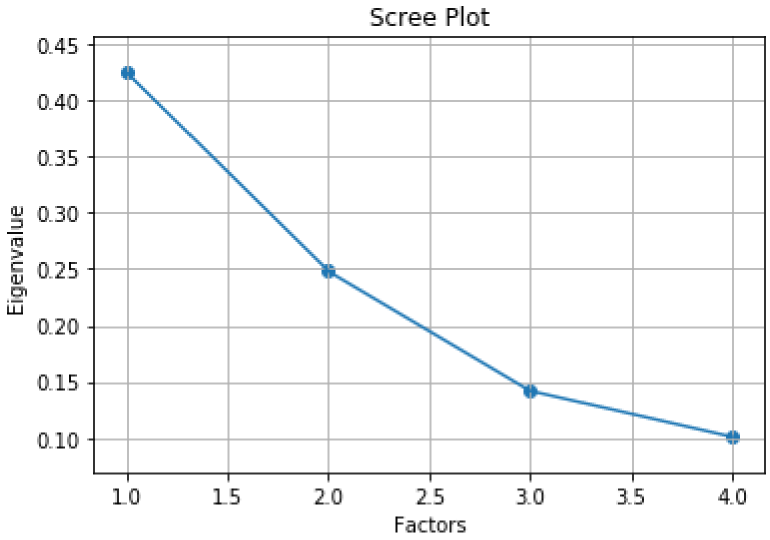
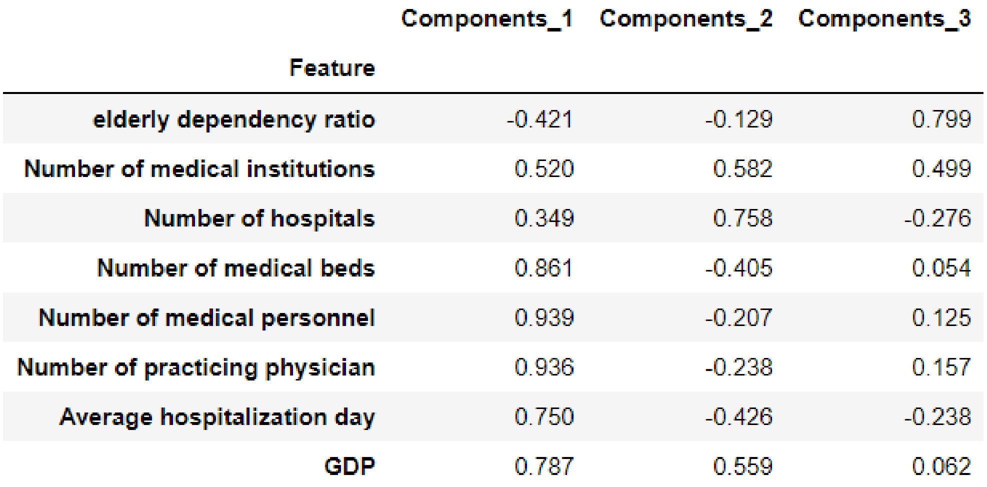


Figure 15: PCA Scree plot



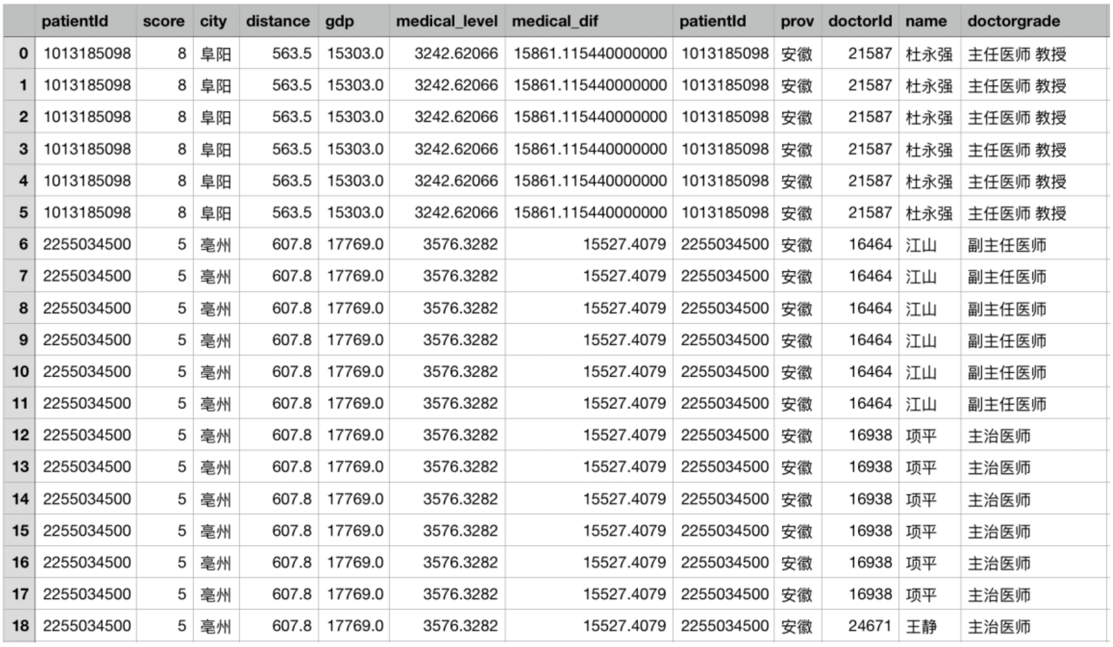
	Components_1	Components_2	Components_3
Feature			
elderly dependency ratio	-0.421	-0.129	0.799
Number of medical institutions	0.520	0.582	0.499
Number of hospitals	0.349	0.758	-0.276
Number of medical beds	0.861	-0.405	0.054
Number of medical personnel	0.939	-0.207	0.125
Number of practicing physician	0.936	-0.238	0.157
Average hospitalization day	0.750	-0.426	-0.238
GDP	0.787	0.559	0.062

Table1: Components matrix

地区	到上海的距离（公里）	人均地区生产总值（全市）/元	医疗水平综合得分	与上海医疗水平的差异
合肥市	408.1	67689	13276.5175	5827.21859
淮北市	532.5	35324	6328.77547	12774.9606
亳州市	607.8	17769	3576.3282	15527.4079
宿州市	503.9	20895	4248.65928	14855.0768
蚌埠市	429.7	35542	6995.21678	12108.5193
阜阳市	563.5	15303	3242.62066	15861.1154
淮南市	451.5	33361	4861.05483	14242.6813
滁州市	315.5	30562	6026.79145	13076.9447
六安市	472.4	19211	4209.93965	14893.7965

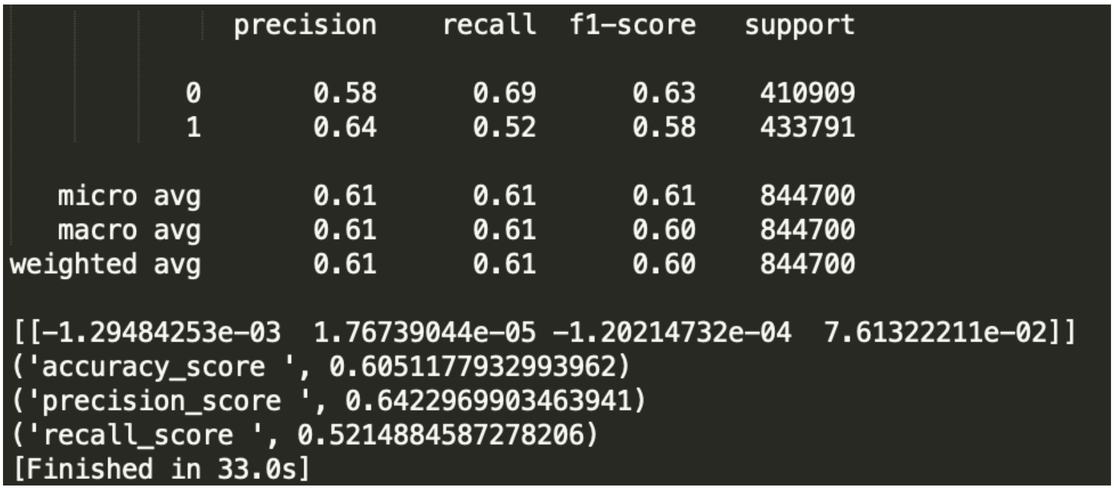
马鞍山市	286.1	60091	10923.092	8180.64415
芜湖市	289.6	64039	12188.9028	6914.8333
宣城市	259.3	35726	6864.70094	12239.0352
铜陵市	350.9	97193	10218.1817	8885.55439
池州市	387	36267	6887.73138	12216.0047
安庆市	424.6	28808	5877.07508	13226.661
黄山市	346.1	37306	7053.7353	12050.0008

Table2: Logistic Regression Data with City Attributes



	patientId	score	city	distance	gdp	medical_level	medical_dif	patientId	prov	doctorId	name	doctorgrade
0	1013185098	8	阜阳	563.5	15303.0	3242.62066	15861.115440000000	1013185098	安徽	21587	杜永强	主任医师 教授
1	1013185098	8	阜阳	563.5	15303.0	3242.62066	15861.115440000000	1013185098	安徽	21587	杜永强	主任医师 教授
2	1013185098	8	阜阳	563.5	15303.0	3242.62066	15861.115440000000	1013185098	安徽	21587	杜永强	主任医师 教授
3	1013185098	8	阜阳	563.5	15303.0	3242.62066	15861.115440000000	1013185098	安徽	21587	杜永强	主任医师 教授
4	1013185098	8	阜阳	563.5	15303.0	3242.62066	15861.115440000000	1013185098	安徽	21587	杜永强	主任医师 教授
5	1013185098	8	阜阳	563.5	15303.0	3242.62066	15861.115440000000	1013185098	安徽	21587	杜永强	主任医师 教授
6	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16464	江山	副主任医师
7	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16464	江山	副主任医师
8	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16464	江山	副主任医师
9	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16464	江山	副主任医师
10	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16464	江山	副主任医师
11	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16464	江山	副主任医师
12	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16938	项平	主治医师
13	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16938	项平	主治医师
14	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16938	项平	主治医师
15	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16938	项平	主治医师
16	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16938	项平	主治医师
17	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	16938	项平	主治医师
18	2255034500	5	亳州	607.8	17769.0	3576.3282	15527.4079	2255034500	安徽	24671	王静	主治医师

Table3: Logistic Regression Data with Patient Attributes



```
             precision    recall  f1-score   support

          0       0.58      0.69      0.63    410909
          1       0.64      0.52      0.58    433791

  micro avg       0.61      0.61      0.61    844700
  macro avg       0.61      0.61      0.60    844700
weighted avg       0.61      0.61      0.60    844700

[[-1.29484253e-03  1.76739044e-05 -1.20214732e-04  7.61322211e-02]]
('accuracy_score ', 0.6051177932993962)
('precision_score ', 0.6422969903463941)
('recall_score ', 0.5214884587278206)
[Finished in 33.0s]
```

Figure 16: Logistic Regression Results

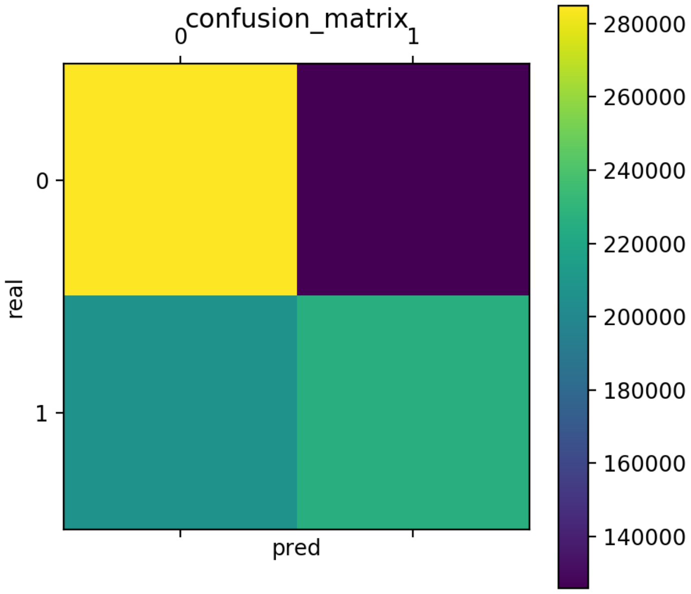


Figure 17: Confusion Matrix