全国人口流动性探究

——基于 200 万酒店入住数据的分析

简介：

在 200w 三字段（姓名，身份证，手机号码），能敏锐的洞察到可以挖掘人口流动信息，为人口政策研究决策提供支持，大大地降低人口普查的成本。

1、 运用 SQL 数据库进行数据清洗、筛选，其次关联表匹配（身份证关联性别、户籍地省市、年龄，手机号码关联运营商，手机号使用地即人口流入地）来对数据源深度分析。

2、 （1）分省外和省内两个维度概括全国流动情况，而省内创新地用 12345 线城市来反映，很好的简化了模型：

利用 12345 线城市各年龄结构来展示城市化进展

利用 12345 线城市之间的流动来刻画城市偏好分析

利用省份间的流动来反映省份、区域的辐射和人口接纳性分析

（2）基于挖掘出的城市吸引力信息，进一步结合多元统计分析探究城市吸引力背后的因素。得出经济和交通对城市魅力作用最大。

（3）结合背景知识和数据分析结果提供个人发展地域选择、针对流入人口多的地区和针对流入人口少流出人口多的地区的建议

3、很好的利用 EXCEL、网页工具地图慧和 powermap 制作、趣化图表，进行客户化数据分析和报告。

目录

1 数据概况

本小组该次进行数据分析的主要对象是某酒店的 200 万条入住客户信息，即“inn200w”。原表中共有三个字段，分别是姓名、身份证号、手机号，其中身份证号、手机号的部分数字被隐藏。共有记录 1970195 条。

应分析需要，我组另搜集到了居民身份证号对应地区、手机号归属地区、中国 1-5 线城市划分、中国省级行政区地域划分、中国各省级行政区 2013 年 GDP 总量及排名、中国各省级行政区二级公路里程数、中国各省级行政区生态文明指数及排名、中国各省级行政区内高校数量、中国各省级行政区人口颜值指数与排名等相关数据以进行综合分析与考量。

2 数据处理

2.1 数据预处理

对于主体对象，即“inn200w”表，因我组研究目标是人口流动性，故首先进行数据去重处理。经相应处理操作后，数据量减少到 1966236 条。

而后，经观察，原记录中存在错误信息，如身份号码、手机号中有出现特殊字符或其他明显不符合要求的情况，所以有必要对上述数据进行进一步的清洗和筛选。

再次，分析过程需要将客户身份证号、手机号与相应地区进行匹配，所以对于某些无法查找到其归属地的信息记录将不得已被剔除。

另外，按照一般情况，对于客户姓名中出现特殊符号等的记录，应视为不合格并剔去，但是考虑到我们本次分析并不涉及姓名的因素，故实际操作中并未舍弃，而是将身份证号、手机号符合要求的记录即视为一个完整的物理人、一条有意义的记录处理分析。

2.2 数据处理思路

本次分析以 SQL Server2008 为基本工具，首先将上述相关表成功导入，并从“inn200w”中提取身份证号、手机号中的隐含信息性别、年龄、出生地、现居地，在此基础上，围绕每一步的中心目标，选择合适的基础表，将其进行关联操作并生成新表，依照各个步骤的数据要求查找并显示结果，清晰明确的数据结果是进行数据分析的首要前提。

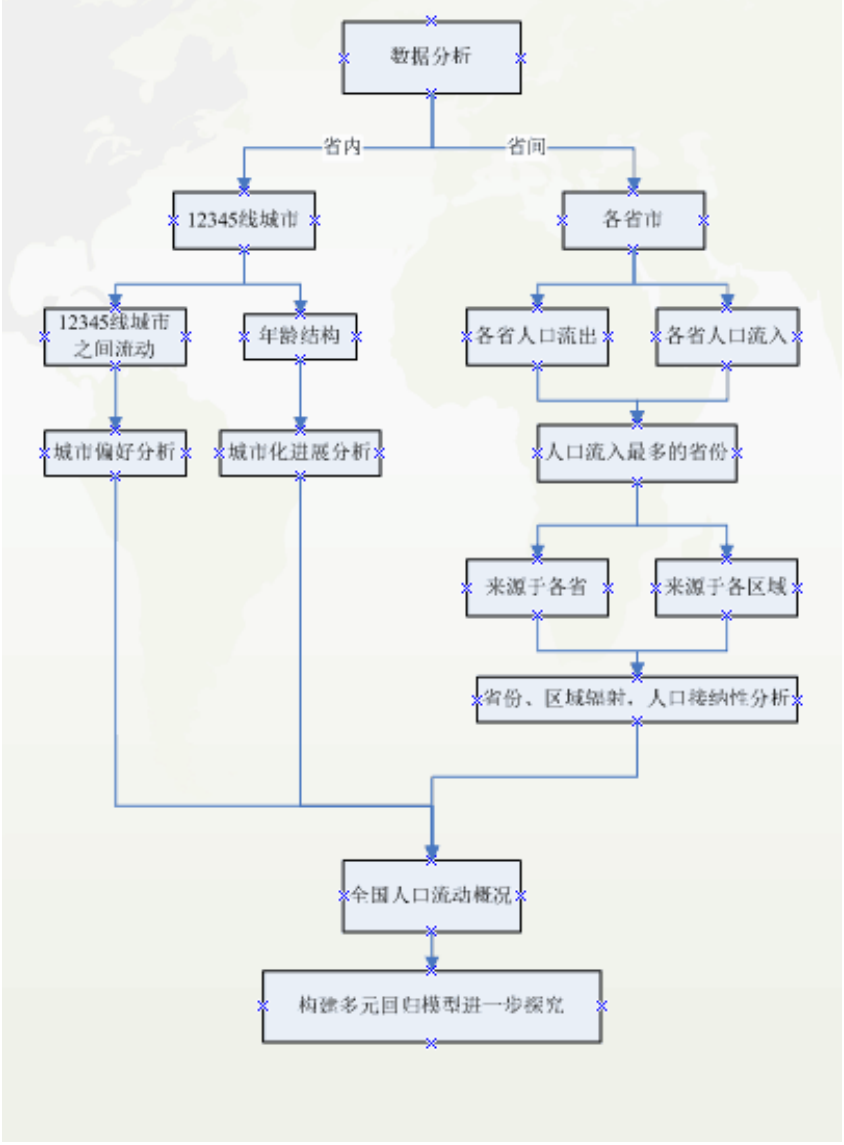
2.3 实现步骤

数据处理的实现步骤顺数据分析的整体思路而走，为各部分分析过程的实现打下基础。为使阅读更加清晰便利，我组将数据处理具体实现的关键 SQL 语句及代码已以批注的形式附在文章中。

3 数据分析

3.1 基本思路

对于经预处理的某酒店 200 万入住用户数据，通过 SQL 的数据表连接可实现将其与身份证号所属地和移动手机地匹配得出其人口籍贯地和现住地信息（代码），故可将这些数据视为在全国范围随机抽取的人口流动数据样本，以对这 200 万数据的研究来反应全国人口流动的情况。



对这代表全国人口流动数据的约 200 万条数据，本文首先将其切割成省内流动和省间流动两个视角进行观察。研究省内人口流动时，若一一将全部省份人口在城市间的流动查询出来，将产生很大的工作量，且得出的数据冗杂不利于整理。因此，本文将通过观察全国所有 12345 线城市之间人口的流动来反应各线城市对外来人口的包容性，通过观察不同年龄阶段的城市间人口流动来折射全国的城市化进程，最终综合反应省内城市流动的常见概况。而研究省间人口流动时，本文将初步研究各省人口流出，流入的情况，随后通过研究人口流入最多的省份，分析该省份流入人口的来源寻得来人口接纳度高的省份，为个人发展地域提供指导和建议。

其次，本文引入颜值、高校数、生态文明指数、GDP、二级公路里程、劳动人口占比为解释变量，该省人口流入量为被解释变量的多元回归模型，进一步探究流入人口最多的省份对人口的吸引力和个人发展的适宜程度。

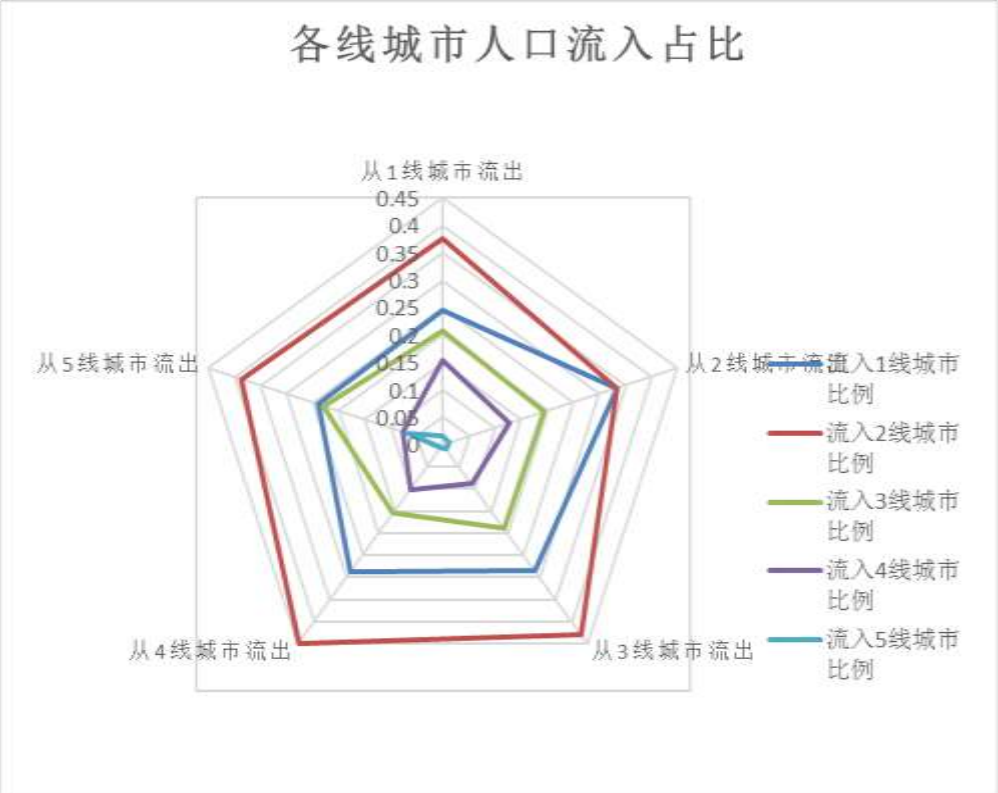
3.2 基于 SQL 数据分析

3.2.1 各线城市之间的人口流动分析

3.2.1.1 各线城市内人口流动分析

通过 SQL 语句可得各线城市之间的流动数据并将其绘制雷达图如下：

	1 线城市流出	从 2 线城市流出	从 3 线城市流出	从 4 线城市流出	从 5 线城市流出
流入 1 线城市比例	0.2459	0.3341	0.2849	0.2871	0.2404
流入 2 线城市比例	0.3752	0.3347	0.4279	0.4478	0.3873
流入 3 线城市比例	0.2079	0.1937	0.1894	0.1546	0.2285
流入 4 线城市比例	0.1550	0.1268	0.0884	0.1026	0.0770
流入 5 线城市比例	0.0160	0.0108	0.0093	0.0078	0.0668



可得以下分析及结论

（1）各线城市人口流动偏好分析

数据：从总体来看，出生于一线城市人口占总人口比例 6.288%，二线 23.0634%，三线 33.0906%，四线 34.8621%。流动后，一线城市人口占总人口比例 29.3361%，二线 40.8952%，三线 18.0487，四线 10.60%。

结论：随着人口的流动，一线，二线城市得到了极大的发展，三线，四线城市稍处于相对不利的地位。

原因：我国的城市化进程依然处于初级中级阶段，正在形成大规模城市圈，所以

人口会集中向一二线城市流动

（2）各线城市对一线城市的偏好分析

数据：所有流动人口中，流动到一线城市的人口占总流动人口比为 29.3362%，将其设为平均水平。所有出生在一线城市又迁居到其他一线城市的人口，占所有出生在一线城市人口数的 24.5946%，低于平均水平。与此同时，出生在二线而流入一线的人口占所有出生在二线的人口 33.4078%，高于平均水平。同理计算出三，四，五线对应比例为 28.4937%，28.7073%，24.0377%，低于平均水平。

结论：在一到五线城市中，二线城市最倾向于流向一线城市。

原因：一线城市对于本来就出生在一线的人来说吸引力不大，人们更倾向于在自己的出生地发展，所以低于平均水平。二线城市出生人口更倾向于流入到一线城市，获得更好的环境。四，五线城市出生人口由于自身所处城市与一线城市差距较大，流入一线城市相对困难，所以低于平均水平。

（3）各线城市偏好二线分析

数据：类似于第一条分析，所有流动人口中，流动到二线城市的人口占总人口比为 40.8954%，视为平均水平。出生在一线而流入二线城市的人口占所有出生在一线城市人口比例为 37.5243%，低于平均水平。同理求出二，三，四，五线城市相应比例为 33.4742%，42.7893%，44.783%，38.7259%。

结论：在一到五线城市中，二线人口最不愿意流向其他二线城市（33.4742%）。三，四线城市出生人口对流入二线城市热情较高。

原因：出生于二线城市的人流入到另一个二线城市好处较小，而三四线出生人口可以获得较大提升。

3.2.1.2 各个年龄段人口流动特征分析

通过 SQL 语句可得各线城市之间的流动数据并将其绘制柱形图如下：

	>60	40-60	18-40	<18
流入 1 线城市比例	0.3788	0.2940	0.2892	0.2610
流入 2 线城市比例	0.3398	0.3759	0.4239	0.3599
流入 3 线城市比例	0.1645	0.1961	0.1758	0.2322
流入 4 线城市比例	0.1078	0.1205	0.1008	0.1363
流入 5 线城市比例	0.0091	0.0136	0.0103	0.0106

得以下分析及结论

（1）一线城市对各年龄段人口吸引力分析

数据：以上图表得知，所有流动人口中，流入一线城市的人口占比为 29%； 60 岁以上人口中，有 37.8759%的人流向了一线；在 40 至 60 岁人口该比为 29.3974%，18 至 40 岁人口中，该比为 28.9211%。小于 18 岁的人口中，该比为 26.1036%。

结论：一线城市对流动人口的吸引力随着年龄的降低而减弱。

原因：我国城市化进程逐步推进，一线城市发展趋于饱和，人们迁入一线城市意愿减弱。

（2）二线城市对各年龄段人口吸引力分析

数据：与上述分析同理，分析二线城市。平均 40%的人口流入二线城市。由图表可知，有 60 岁以上的人有 33%流入二线，在 40 至 60 岁人口中，有 37%，18 至 40 岁人口中，有 42%。小于 18 岁的人口中，有 35%。

结论：在各个年龄人口中，18 至 40 岁的人口流入二线城市的积极性最高。

原因：处于这个年龄段的人，在迁居的时候，正值一线城市发展速度放慢，城市化进程辐射到二线城市的时候，所以他们更倾向于迁入二线城市。

3.2.2 省间人口流动情况分析

3.2.2.1 省间人口流动概况

通过 SQL 语句可得各线城市之间的流动数据并将其在地图中可视化如下：

pro1	人口流入占比	人口流出占比
江苏	0.078279565	0.577660506
上海	0.070071179	0.242690971
广东	0.056357246	0.682851631
北京	0.051912512	0.340475826
山东	0.047504359	0.512886341
浙江	0.035102652	0.516633893

河南	0.021925222	0.610812831
安徽	0.018757297	0.754025662
福建	0.018238793	0.604229078
湖北	0.01823531	0.679026552
河北	0.018178988	0.611584966
辽宁	0.018013508	0.572142257
山西	0.017964155	0.604614794
陕西	0.017452619	0.598949849
四川	0.013936315	0.790065014
黑龙江	0.011358311	0.686140437
天津	0.011019222	0.359587274
内蒙古	0.009154234	0.752187325
湖南	0.008953336	0.803245612
吉林	0.007216087	0.695534135
江西	0.006896159	0.732741598
贵州	0.006651713	0.745144202
重庆	0.006222046	0.572870077
广西	0.005906763	0.643158128
云南	0.005848119	0.751732717
新疆	0.005767992	0.80594684
甘肃	0.004762919	0.731103221
海南	0.003504691	0.870923603
宁夏	0.001794731	0.665137615
青海	0.001548543	0.752504238
西藏	0.000531858	0.883173496

（1）各省份流出人口分析

结论：由上图和数据可以看出西部地区、内蒙古、湖南、安徽流出人口占本省人口比重大

原因：西部地区、内蒙古经济欠发达，交通不便，贫瘠土地多不适宜居住，造成大量人口流出。而湖南、安徽地处中原，毗邻广东、江西、浙江等经济大省，人口易迁移。

（2）各省份流入人口分析

结论：东南沿海流入人口占全部流动人口比例最高，其中外来人口流入华东区域最多。

原因：东南沿海经济发达，全国人口流入。

3.2.2.2 探究人口流入最多的前六个省份人口来源

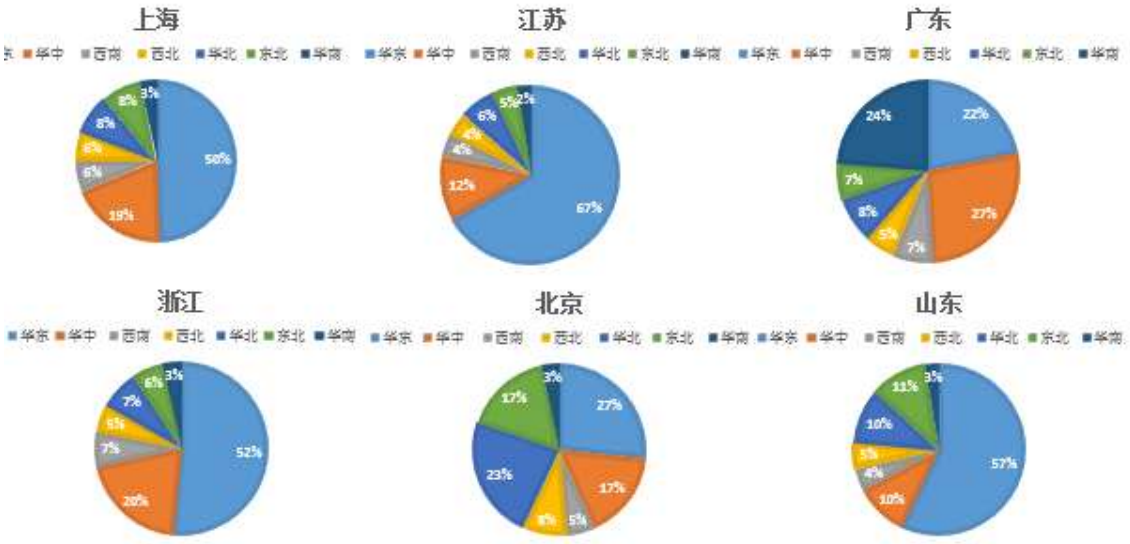
步骤一：通过 SQL 语句获取该六个省级行政区流入人口在全国各省的来源比例

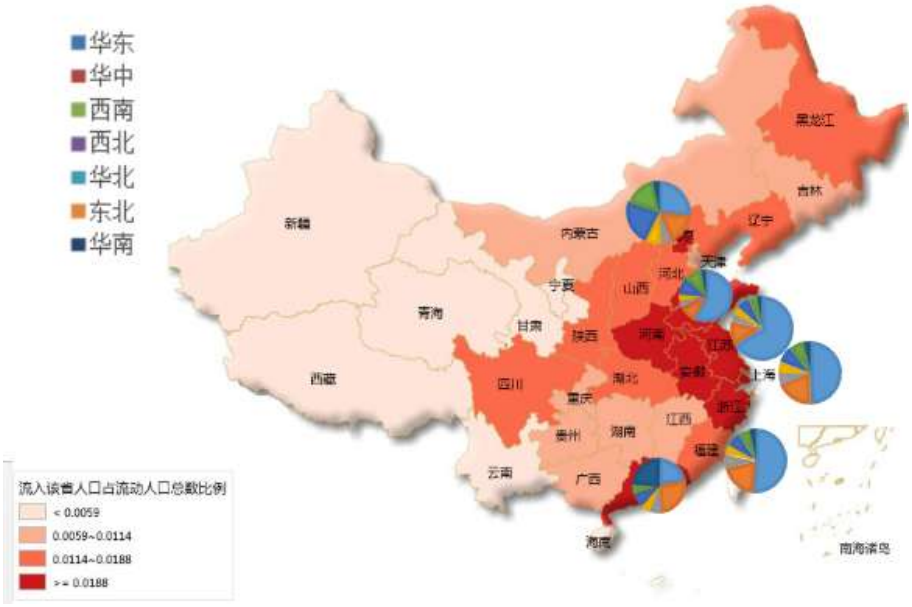
人口来源省份	江苏	上海	广东	北京	山东	浙江
江苏	0.451883	0.187817	0.062022	0.063798	0.077076	0.088345
上海	0.028757	0	0.022563	0.023533	0.022038	0.03075
广东	0.015933	0.021138	0.203664	0.021083	0.01941	0.023356
北京	0.010978	0.016879	0.013919	0	0.016024	0.011132
山东	0.042568	0.056711	0.036739	0.082779	0.388547	0.044793
浙江	0.043941	0.094729	0.034359	0.033219	0.036045	0.253209
河南	0.03745	0.053629	0.049494	0.067847	0.040873	0.050781
安徽	0.074819	0.107871	0.034586	0.044191	0.029811	0.07235
福建	0.024974	0.048069	0.028487	0.018858	0.01831	0.026581
湖北	0.032006	0.051077	0.077837	0.042558	0.025937	0.051591
河北	0.01821	0.02195	0.023253	0.113011	0.032757	0.025407
辽宁	0.017691	0.029748	0.024623	0.061539	0.028968	0.021321
山西	0.015755	0.019473	0.022429	0.053262	0.024421	0.018857
陕西	0.018447	0.024934	0.02928	0.040836	0.023138	0.028235
四川	0.022148	0.034272	0.037378	0.030311	0.018554	0.033429
黑龙江	0.017134	0.0293	0.023356	0.062333	0.046275	0.023141
天津	0.008241	0.009596	0.008531	0.023119	0.010328	0.00708
内蒙古	0.009776	0.012678	0.012961	0.042726	0.018652	0.010669
湖南	0.019664	0.033816	0.084121	0.030579	0.017906	0.037267
吉林	0.013552	0.020666	0.017628	0.042614	0.032023	0.016888
江西	0.026851	0.053057	0.059951	0.025591	0.016427	0.060093
贵州	0.006883	0.008783	0.01327	0.007684	0.007383	0.011049
重庆	0.006728	0.013101	0.012301	0.009239	0.006832	0.010024
广西	0.005059	0.008187	0.025932	0.007684	0.005659	0.007873
云南	0.005474	0.006646	0.007171	0.006174	0.006185	0.01009
新疆	0.007373	0.012786	0.008644	0.014775	0.009693	0.007443
甘肃	0.010829	0.014103	0.011467	0.017773	0.010805	0.011116
海南	0.001298	0.002411	0.007315	0.002528	0.001283	0.001439
宁夏	0.00184	0.003331	0.002637	0.00529	0.002787	0.002283
青海	0.003264	0.00285	0.002916	0.004463	0.00429	0.002845
西藏	0.000475	0.000864	0.001164	0.000604	0.001565	0.000562

由于各省流入这六个省份的人口比例比较低，且数据差异小，不易看出，故进行步骤二，将分析六个省份来自各省份的人口转为对于六个省份来自各区域的人口分析。

步骤二通过 SQL 语句获取全国各区域流入六个省份人口比例

流入省份	华东	华中	西南	西北	华北	东北	华南
江苏	0.6669	0.1160	0.0417	0.0418	0.0630	0.0484	0.0223
上海	0.4952	0.1916	0.0632	0.0580	0.0806	0.0797	0.0317
广东	0.2188	0.2714	0.0713	0.0549	0.0811	0.0656	0.2369
北京	0.2664	0.1666	0.0540	0.0831	0.2321	0.1665	0.0313
山东	0.5718	0.1011	0.0405	0.0507	0.1022	0.1073	0.0264
浙江	0.5160	0.1997	0.0652	0.0519	0.0731	0.0614	0.0327





区域	出生于各大区人口数	出生人数占比
华北	142583	0.1382
西南	71523	0.0693
东北	105066	0.1018
华南	55064	0.0534
华东	387599	0.3757
西北	77084	0.0747
华中	192831	0.1869

（1）华东区域流动分析

数据：从图中圆饼图和全国各区域流入六个省份人口比例图可以看出，人口流入前六个省份就有江苏、上海、山东、浙江四个来自华东地区。而这四个省份的流入人口主要来源于华东地区，比例平均超过 50%；华东地区流入其他两个省份广东、北京与其他区域流入的数量相差不大；江苏流入人口中来自华东地区最多高达 66.9%.

结论：华东区域对其本地域的人口流入接纳性较高，而华东地区的人口流出除华东地区外也泛及华北、华南地区。虽然浙江与江苏在地理位置，经济情况有许多的相似性，但浙江与江苏相比，江苏对于华东更有吸引力。

原因：华东人多，经济发达，对全国辐射能力强，对自身区域的外来人口接纳度较高。

（2）华北、华南区域流动分析

数据：在华北、华南区域均只有一个城市在流入人口最多的前六个省份中，而这两个省份北京、广东来自本区域在全部流入人口的占比只有 0.23、0.24，和来自华东区域的比例相差不大。

结论：北京、广东作为华北、华南的经济中心，起着非常重要的辐射带动作用，对全国各地的人口包容性也较大。

原因：北京作为首都，在华北整体经济衰落的大背景下，成为华北流动人口向往的地区。华南地区广东作为经济大省，其他省份与之差距较大。

综上可得出，在区域发展中，华东区域总体发展较快且均匀、区域间交流多；而华北、华南地区地区发展差距大，出现了像北京和广东一枝独秀的现象，因此应加快华北地区、华南地区的城市化进程、减少发展差距。

3.3 基于 eview 多元回归分析

3.3.1 基本思路

为探究各省适合个人发展的因子和省份本身的吸引力，本文从审美、教育、生态环境、经济发展、交通便利、该地劳动力情况选择颜值、高校数、生态文明指数、13GDP 二级公路里程、劳动人口占比颜值，生态文明指数，劳动力人口作为解释变量，以该省人口流入量为被解释变量，通过 eview 构建多元回归模型探究各解释变量对该省人口流入量的影响。

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

其中X_1、X_2、X_3、X_4、X_5、X_6为颜值、高校数、生态文明指数、13GDP 二级公路里程、劳动人口占比颜值，生态文明指数，劳动力人口；Y为该省人口流入量

3.3.2 样本说明

通过统计年鉴和百度截取 31 个省份样本数据如下

pro1	颜值	高校数	生态文明指数	13GDP	二级公路里程	劳动人口占比	流入该省人口总量
江苏	2.969	134	79.11	27603.98	21779	0.653970572	80764.94119
上海	2.779	66	82.58	10168.52	3069	0.639397618	72295.93893
广东	2.931	130	86.23	28465.92	19044	0.704394891	58146.58856
北京	2.862	84	92.11	9112.8	3279	0.663961854	53560.73426
山东	2.917	130	76.71	25958.2	24151	0.68077573	49012.6224
浙江	3.309	82	91.57	16953.9	9224	0.683136159	36217.1612
河南	2.774	121	71.95	14556.63	24981	0.645185482	22621.3478
安徽	4.33	107	74.41	8591.3	10640	0.651943669	19352.84118
福建	3.604	79	86.56	8855.26	7510	0.719390567	18817.87468
湖北	2	99	74.59	10949.3	17135	0.660639295	18814.28109
河北	3.233	101	65.85	13154.6	16728	0.705837492	18756.17087
辽宁	3.2	104	90.64	12334.7	17250	0.614419442	18585.43688
山西	2.227	71	76.66	6016.6	5587	0.658138251	18534.51692
陕西	2.085	80	76.55	6777.73	7611	0.66986069	18006.73965
四川	2.614	97	87.05	11655.1	13140	0.65886122	14378.793
黑龙江	3.5	78	88.17	5545.1	8849	0.607668253	11718.93737
天津	3.33	45	79.62	6579.01	3244	0.655034842	11369.0823
内蒙古	2.959	48	84.38	7087.55	13689	0.576067959	9444.88093
湖南	3	109	85.92	10921.8	9406	0.660133121	9237.604418

吉林	2.8	52	80.91	4808.01	8756	0.635769917	7445.197762
江西	4.33	82	88.6	5901.63	9464	0.61646913	7115.112048
贵州	2.375	47	80.83	3249.85	3831	0.659318383	6862.904888
重庆	3.125	57	90.11	5840.51	7522	0.687712298	6419.595961
广西	2.833	61	85.4	5810.18	9132	0.683244024	6094.302725
云南	1.966	60	83.53	4640.59	9553	0.690035753	6033.796778
新疆	4.5	39	78.53	3352.09	11099	0.677322031	5951.125746
甘肃	4	38	75.95	2349.57	5856	0.603699358	4914.141678
海南	3.333	109	93.27	1516.69	1359	0.65995153	3615.964939
宁夏	2.8	16	69.38	1008.16	2567	0.616756865	1851.713709
青海	3.056	11	80.23	877.349	5289	0.587409314	1597.70924
西藏	4.3	6	88.53	329.59	9556	0.682857143	548.7444915

在三维地图数据概况如下

3.3.3 构建多元模型

设：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

其中X_1、X_2、X_3、X_4、X_5、X_6为颜值、高校数、生态文明指数、13GDP 二级公路里程、劳动人口占比颜值，生态文明指数，劳动力人口；Y为该省人口流入量

步骤一：将数据导入 eview，通过最小二乘法（OLS）让 Y 对X_1、X_2、X_3、X_4、X_5、X_6进

行多元回归。得到结果如下：

Dependent Variable: Y
Method: Least Squares
Date: 12/06/15 Time: 11:52
Sample: 1 31
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	63668.88	50530.00	1.260021	0.2198
X1	-263.7942	3306.489	-0.079781	0.9371
X2	-35.16929	101.3320	-0.347070	0.7316
X3	-146.1699	333.1121	-0.438801	0.6647
X4	3.684955	0.569152	6.474463	0.0000
X5	-1.842439	0.558944	-3.296287	0.0030
X6	-65044.39	65587.01	-0.991727	0.3312

R-squared	0.757133	Mean dependent var	19938.28
Adjusted R-squared	0.696416	S.D. dependent var	21090.15
S.E. of regression	11620.34	Akaike info criterion	21.75458
Sum squared resid	3.24E+09	Schwarz criterion	22.07839
Log likelihood	-330.1960	Hannan-Quinn criter.	21.86013
F-statistic	12.46992	Durbin-Watson stat	1.967874
Prob(F-statistic)	0.000002		

从t检验及其伴随概率来看，只有变量X_4、X_5较为显著，其他解释变量均不显著；并且方程拟和优度$\overline{R^2}$为 0.6964、同时方程整体 F 值 12.46992>F（6,24）F 检验很显著。因此可以怀疑在变量其他解释变量之间存在多重共线性。

步骤二：修正多重共线性

首先对X_1、X_2、X_3、X_4、X_5、X_6绘制自相关系数图：

	X1	X2	X3	X4	X5	X6
X1	1.000000	-0.167945	0.162804	-0.171936	-0.068361	-0.086815
X2	-0.167945	1.000000	0.024315	0.787322	0.636325	0.308385
X3	0.162804	0.024315	1.000000	-0.051022	-0.258814	0.046321
X4	-0.171936	0.787322	-0.051022	1.000000	0.751122	0.328133
X5	-0.068361	0.636325	-0.258814	0.751122	1.000000	0.152037
X6	-0.086815	0.308385	0.046321	0.328133	0.152037	1.000000

可看出X_2、X_4、X_5间高度相关，原模型存在多重共线性，故采用逐步回归法，让

让 Y 对X_1、X_2、X_3、X_4、X_5、X_6分别进行一元回归。

P 值	R 方
0.3317	0.000
0.4388	0.2997
0.9091	-0.0340
0.0000	0.6147
0.0396	0.1082
0.2554	0.0114

在 5%的置信水平下，拟合度达 0.6 以上的只有X_4变量，但考虑到X_2、X_4、X_5间高度相关，故将 Y 对X_2、X_4、X_5中选取两个进行二元回归。

发现在 5%的置信水平下，Y 对X_4、X_5进行二元回归时拟合程度提升到 0.7237

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6976.697	3854.124	1.810190	0.0810
X4	3.376081	0.416806	8.099891	0.0000
X5	1.707074	0.483800	3.528471	0.0015
R-squared	0.742158	Mean dependent var		19938.28
Adjusted R-squared	0.723740	S.D. dependent var		21090.15
S.E. of regression	11085.07	Akaike info criterion		21.55635
Sum squared resid	3.44E+09	Schwarz criterion		21.69512
Log likelihood	-331.1235	Hannan-Quinn criter.		21.60159
F-statistic	40.29672	Durbin-Watson stat		2.010844
Prob(F-statistic)	0.000000			

故最终得到模型为

$$Y=6976.697+3.376081X_4+1.707074X_5$$

（0.1810）　（8.100）　（3.528）

模型表明颜值，生态文明指数，劳动力人口，高校数均不影响人口流动；GDP，二级公路正向影响流入该省人口数；高校数、生态文明指数和二级公路里程三个变量之间高度相关。

3．4 汇总结论：

从代表省内城市流动的全国 12345 线城市之间流动可看出 2 线城市人口流入最多，在城市偏好研究中有二线城市最倾向于流向一线城市、三，四线城市出生人口对流入二线城市热情较高等现象；而在研究人口流动的年龄结构时，可发现在各个年龄人口中，18 至 40 岁的人口流入二线城市的积极性最高，且随着年龄段的降低，人口流入 345 线城市的比例升高体现出中国的城市化进程成果显著。

而从省间人口流动来看西部地区、内蒙古、湖南和安徽人口流出情况较为严重、东南沿

海经济发达，全国人口流入；华东区域内部人口流动最为频繁，对其本地域的人口流入接纳性较高，而华东地区的人口流出除华东地区外也泛及华北、华南地区；华北华南地区主要以北京和广东为辐射中心，对全国各地的人口包容性也较大。

最后在进一步探究流入人口最多的省份对人口的吸引力和个人发展的适宜程度时，引入了多元回归分析。分析表明，代表经济状况的 GDP 和交通便利程度的二级公路显著影响流入该省的人口数，而代表教育的高校数变量与 GDP 高度相关，说明高校数的增多是 GDP 增高带来的产物。而模型中代表审美的颜值，代表环境的生态文明指数和劳动力人口比例与流入该省的人口数显著不相关，分别说明了各省份的审美并不兼容，非自己家乡的美女数量对自己流入该省的吸引力不大；人们在省份间流动时对环境的注重程度不大，中国对民众的环保意识教育有待加强；劳动力流入比例在各省分布均匀且均超 0.5，体现了中国省份对 18-40 年龄阶段的劳动力人口接纳性均高。

4 结论与建议

4.1 建议提出方案

4.1.1 个人发展地域选择

人口流入多的地区可以表明该地较适宜发展，基于该结论提出以下关于个人发展地域最佳选择方案：

个人发展地域选择方案			
省内	所在城市	2 线城市	3、4、5 线城市
	最佳流向城市	1 线城市	2 线城市
省间	所在区域	华东	华东以外
	最佳流向区域	华东	华北、华南
最佳外出发展年龄段		18-40	
流向城市偏好选择顺序：		2、1、3、4、5 线城市	

4.1.2 针对流入人口多的地区的建议：

（1）一线城市发展速度放慢，二线城市迎来发展的高峰期，国家应顺势引导，大力发展二线城市。

（2）流入人口最多的几个省份在各自大区起着明显的带动作用，应加强各省带动作用。

4.1.3 针对流入人口少流出人口多的地区的建议

（1）当地政府应加快经济建设努力增强城市吸引力，合理规划健全城市人口流失预警机制，以防出现劳动力流失、留守儿童老人增多等问题而激化社会矛盾。

（2）面对当前网络上种种关于 GDP 无用论，嘲讽道路修建的论调，应加以引导，继续坚持以经济建设为中心，加强基础设施建设。西北人口向东南沿海流入趋势明显，地区发展不平衡。国家应继续大力支持西部大开发、减少区域差距。

（3）三四线城市应加强与二线城市的联通与交流，在道路修建，政策优惠时更多考虑三四线城市。

4.2 不足与反思

（1）处理数据的过程中，有时会产生数据丢失的情况，一定程度上影响分析结果的精确

（2）虽然小组成员都尽了自己最大的努力，但囿于时间、能力限制，分析与结论较为浅薄，有待在以后进一步的学习中加强提高。

处理匹配身份信息时，有些行政区域划分产生过变化，某些旧身份证前 4 位所代表的地区现已无法匹配，导致样本数目的减少。

（3）多元回归分析中，因研究对象较为特殊，在相互独立的影响变量寻找上，还需要

进一步深入探究完善。

（4）数据来源不明确，可信度未知，时间跨度也不清楚，一定程度上影响我们的分析过程及结果，比如其中不少以当前时间计算年龄超过 100 岁的住户数据，今后在目标数据的选择上应多加注意。