侯正罡　2018140713

2018 年 11 月 28 日

实验一任务说明：

1. 熟悉 SPSS 和 EXCEL 等数据分析软件的使用；
2. 基于你的理解，给出给定数据文件的描述报告；
3. 识别给定数据文件中的连续属性，并采用你认为合适的方法完成连续属性的离散化；

数据描述:

行数：1411

列数：35

非数值类型：产品 ID，型号，品牌

连续数值类型：颜色数，上市时间，芯片主频，频段数量，零售价格，厚度，屏幕数量，产品重量，屏幕尺寸，分辨率，RAM，ROM，Flash 内存，摄像头，电池容量，文字输入方法数

非连续数值类型：市场定位，芯片平台，AP，触摸屏， 键盘类型，外观类型，定位
FM 广播，电视，Modem，红外，蓝牙，WLAN，重力感应器，方向感应器，智能系统

给出连续数值类型报告表：

类别	最大值	最小值	平均值	方差
颜色数	7	1	1.542877	0.80862122
上市时间	2012	2003	2009.607	2.16077828
芯片主频	2400	40	196.146	43967.2567
频段数量	5	1	1.647768	1.09499324
零售价格	9380	184	1117.02	1101455.25
厚度	85	9	14.94894	19.9044645
屏幕数量	2	1	1.047484	0.0452614
产品重量	790.2	48.4	107.603	2026.05207
屏幕尺寸	7	0	2.411389	0.47865991
分辨率	921600	6240	82613.7	7597228646
RAM	4096	1	139.4232	75964.4873
ROM	16384	0	306.5259	1355748.93
Flash 内存	16384	0	239.5981	1861591.54
摄像头	1300	0	125.4784	23084.0795
电池容量	4000	100	1134.894	138075.485
文字输入方法数	3	1	1.647059	0.63988319

数据分析：

给出

将连续数值数据按照时间分类求平均

给出 SPSS 分析报告：

报告

上市时间		颜色数	芯片主频	频段数量	零售价格	厚度	产品重量	屏幕数量	屏幕尺寸	分辨率	RAM	ROM	Flash内存	摄像头	电池容量	文字输入方法数
2003	平均值	1.00	68.00	1.00	1689.00	20.2500	99.0000	1.00	2.2000	48640.00	128.000	320.00	22.5000	80.00	1005.00	2.00
	个案数	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	标准偏差	0.000	39.598	0.000	1825.750	2.05061	15.55635	0.000	0.28284	39824.254	0.0000	271.529	10.60660	70.711	63.640	1.414
2004	平均值	1.00	147.00	1.00	1000.00	24.3000	110.000	1.00	3.2000	96000.00	128.000	256.00	140.0000	200.00	1100.00	2.00
	个案数	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	标准偏差															
2005	平均值	1.00	85.36	1.00	700.45	17.7000	100.5818	1.00	2.0873	65250.91	89.455	105.82	64.7091	90.00	1089.09	1.27
	个案数	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
	标准偏差	0.000	32.809	0.000	583.807	3.55837	21.00294	0.000	0.52637	56278.067	109.1351	106.017	101.84068	126.570	209.067	0.647

侯正罡　2018140713

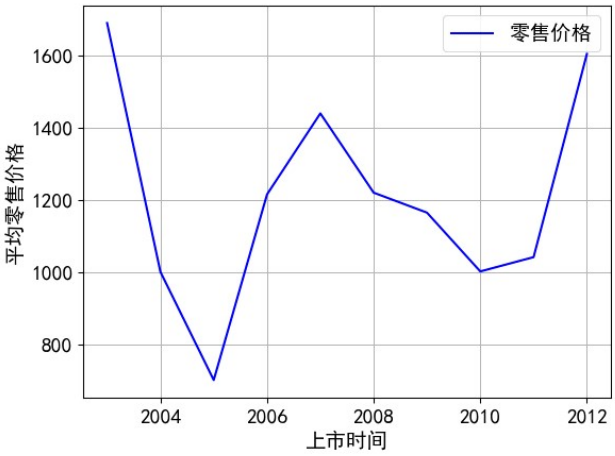
2006	平均值	1.00	96.74	1.36	1214.51	18.2385	105.2000	1.03	2.4185	86302.95	109.475	186.10	261.0656	159.18	1139.84	1.82
	个案数	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61
	标准偏差	0.000	37.624	0.775	905.687	3.70858	25.65700	0.180	0.68035	85028.392	123.2698	188.606	1098.28954	171.962	308.774	0.847
2007	平均值	1.12	95.97	1.67	1438.87	16.1925	100.2817	1.02	2.3263	72747.20	128.000	186.13	81.4483	114.17	1047.17	1.60
	个案数	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60
	标准偏差	0.372	31.043	1.084	1391.596	3.38687	25.49170	0.129	0.71045	72068.952	177.4791	210.658	179.95201	143.837	358.958	0.807
2008	平均值	1.20	107.66	1.57	1219.32	16.1832	100.2079	1.09	2.3391	71315.96	108.168	273.41	363.1545	135.84	1063.76	1.72
	个案数	101	101	101	101	101	101	101	101	101	101	101	101	101	101	101
	标准偏差	0.510	68.634	0.952	1187.583	4.49317	22.10868	0.286	0.59319	65657.296	123.2458	902.026	1447.77167	142.452	313.499	0.814
2009	平均值	1.17	122.69	1.45	1164.03	14.9489	103.6151	1.05	2.3588	76280.24	127.607	355.90	247.7807	122.93	1093.82	1.59
	个案数	338	338	338	338	338	338	338	338	338	338	338	338	338	338	338
	标准偏差	0.547	78.772	0.904	1044.702	3.14220	30.70446	0.225	0.67721	84551.414	273.5810	1577.986	1487.88354	151.160	330.573	0.796
2010	平均值	1.64	163.32	1.73	1001.28	14.6808	110.1327	1.04	2.4225	85686.71	135.105	341.14	253.8085	122.44	1166.48	1.67
	个案数	410	410	410	410	410	410	410	410	410	410	410	410	410	410	410

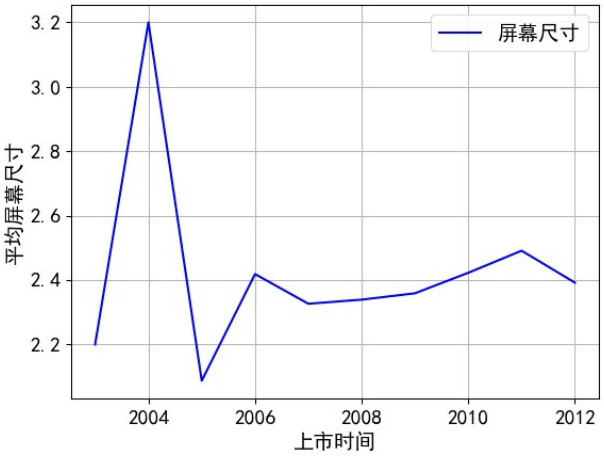
侯正罡　2018140713

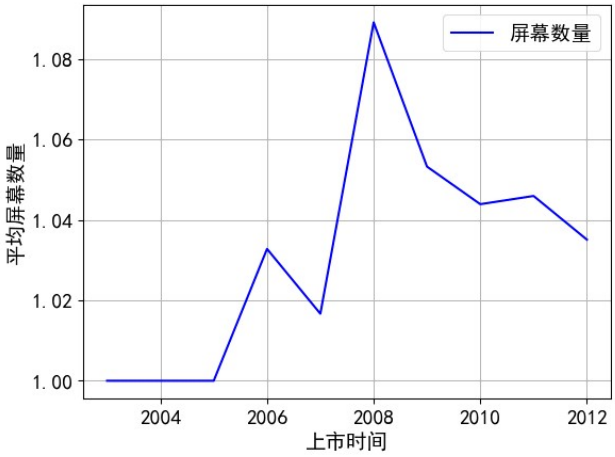
	标准偏差	0.928	132.647	1.065	1009.954	4.82366	53.42931	0.205	0.68725	97851.261	258.4841	1437.010	1473.01417	164.770	417.303	0.810
2011	平均值	1.99	306.15	1.78	1040.80	14.2219	111.5022	1.05	2.4911	87012.06	163.684	275.47	140.9681	122.78	1177.54	1.61
	个案数	370	370	370	370	370	370	370	370	370	370	370	370	370	370	370
	标准偏差	1.050	273.830	1.137	980.072	5.11969	52.65001	0.210	0.71732	81018.347	343.3292	566.113	924.63120	135.791	387.217	0.772
2012	平均值	1.93	549.18	1.93	1603.58	13.6995	112.7414	1.04	2.3919	100279.58	192.772	319.79	696.7742	144.56	1101.58	1.79
	个案数	57	57	57	57	57	57	57	57	57	57	57	57	57	57	57
	标准偏差	1.083	387.287	1.280	1118.558	2.83524	58.59821	0.186	0.79665	111190.855	328.1911	587.426	2580.01073	169.570	297.884	0.840
总计	平均值	1.54	196.15	1.65	1117.02	14.9489	107.6030	1.05	2.4114	82613.70	139.423	306.53	239.5981	125.48	1134.89	1.65
	个案数	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411	1411
	标准偏差	0.899	209.684	1.046	1049.502	4.46144	45.01169	0.213	0.69185	87162.083	275.6166	1164.366	1364.40153	151.934	371.585	0.800

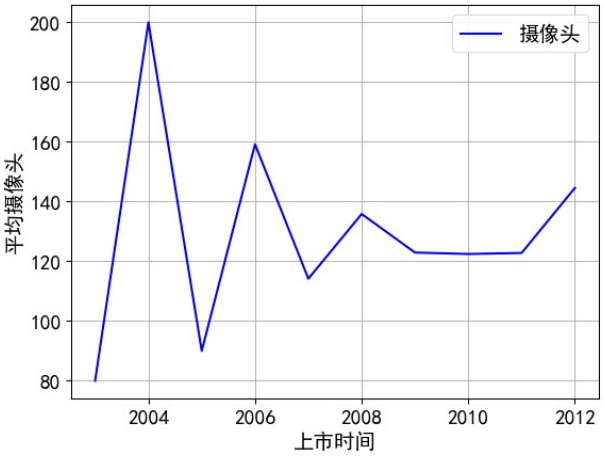
给出观察图：

无规律：

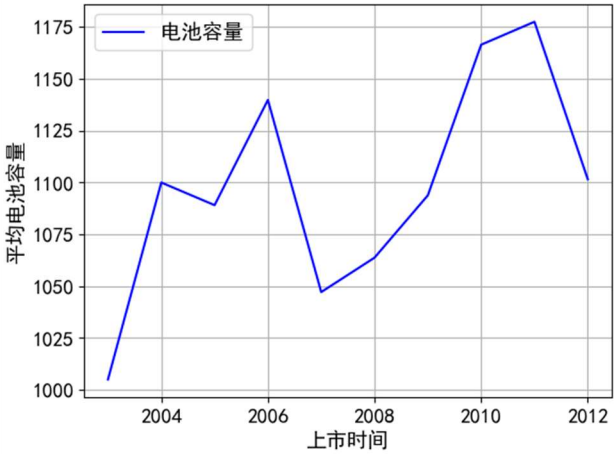

零售价格
平均零售价格
1600
1400
1200
1000
800
2004
2006
2008
2010
2012
上市时间


屏幕尺寸
平均屏幕尺寸
3.2
3.0
2.8
2.6
2.4
2.2
2004
2006
2008
2010
2012
上市时间

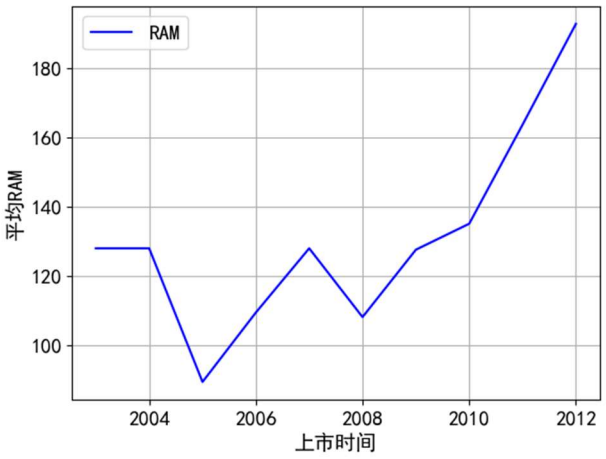

屏幕数量
平均屏幕数量
1.08
1.06
1.04
1.02
1.00
2004
2006
2008
2010
2012
上市时间

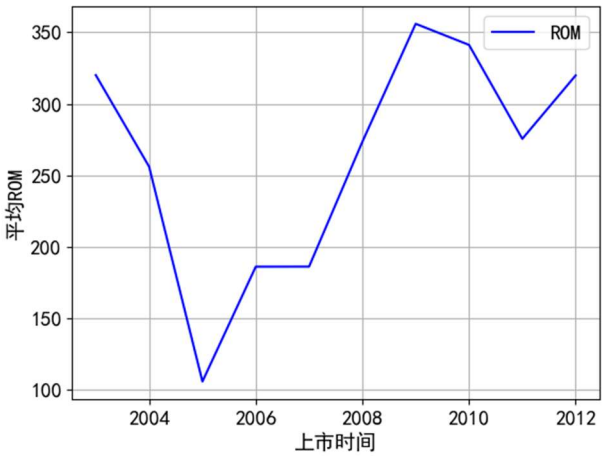

摄像头
平均摄像头
200
180
160
140
120
100
80
2004
2006
2008
2010
2012
上市时间


文字输入方法数
平均文字输入方法数
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
2004
2006
2008
2010
2012
上市时间


电池容量
平均电池容量
1175
1150
1125
1100
1075
1050
1025
1000
2004
2006
2008
2010
2012
上市时间

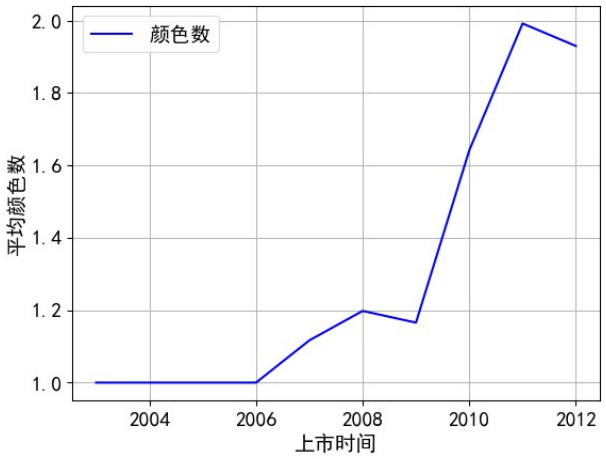
上升态：

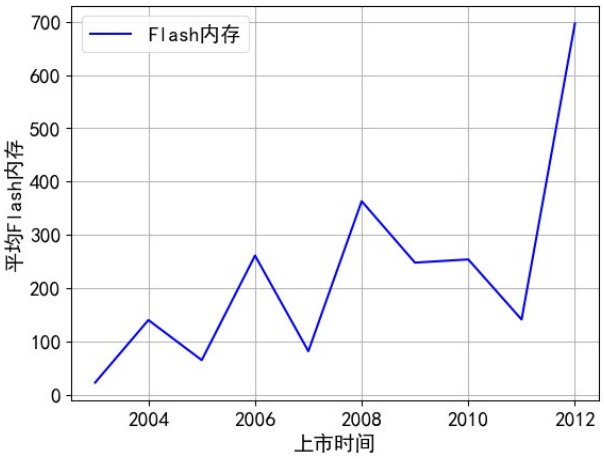

RAM
平均RAM
上市时间
100
120
140
160
180
2004
2006
2008
2010
2012

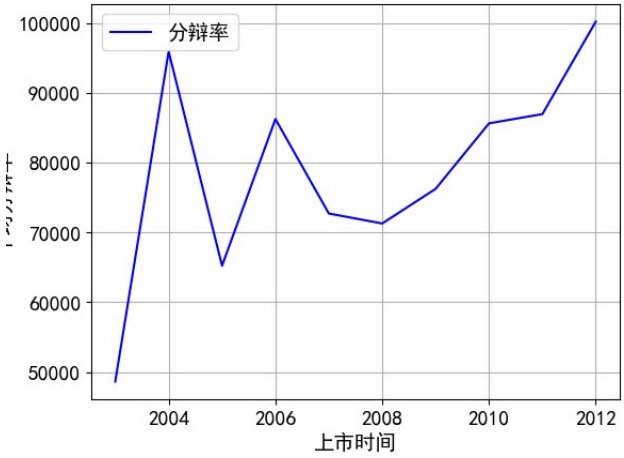

ROM
平均ROM
上市时间
100
150
200
250
300
350
2004
2006
2008
2010
2012


频段数量
平均频段数量
上市时间
1.0
1.2
1.4
1.6
1.8
2004
2006
2008
2010
2012

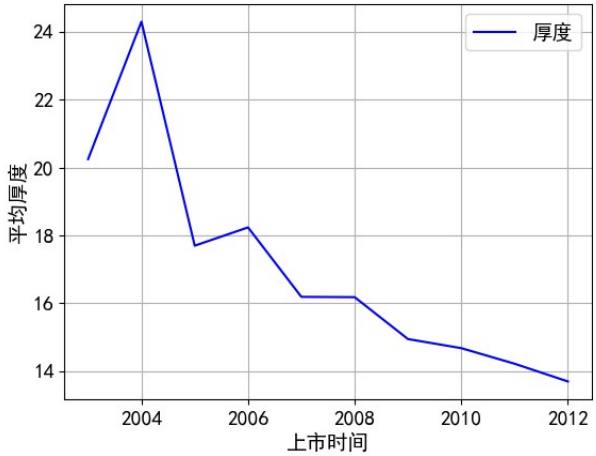

芯片主频
平均芯片主频
上市时间
100
200
300
400
500
2004
2006
2008
2010
2012


颜色数
平均颜色数
上市时间
1.0
1.2
1.4
1.6
1.8
2.0
2004
2006
2008
2010
2012


Flash内存
平均Flash内存
上市时间
0
100
200
300
400
500
600
700
2004
2006
2008
2010
2012



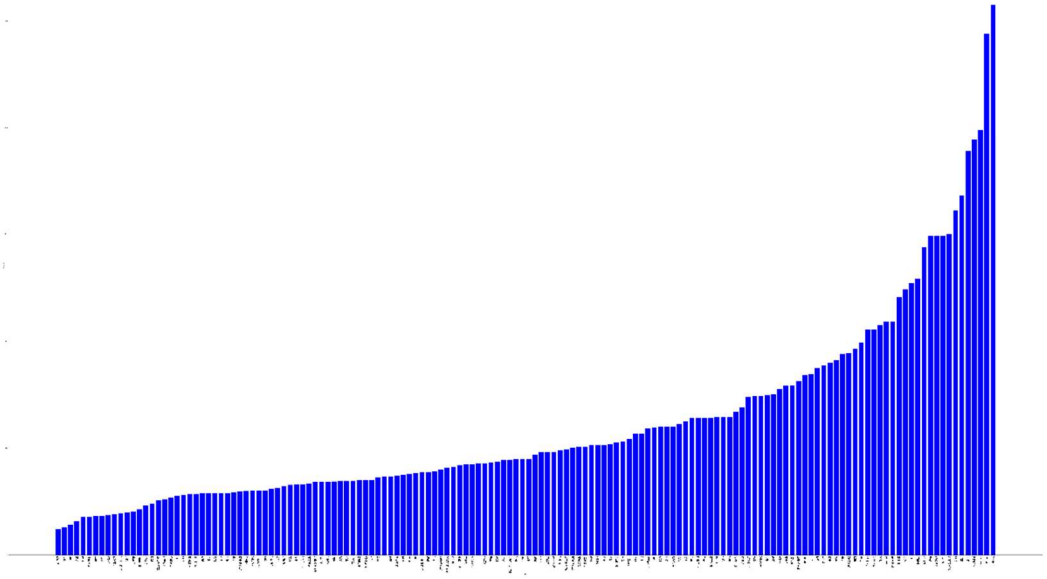
下降态：



经观察：

1. 芯片主频，频段数量，厚度，产品重量，分辨率，RAM，ROM，Flash 内存，颜色数程上升态；厚度程下降态；其他连续数据无明显规律。
2. 在 2005 年众多数据出现峰值，该年份数据十分有可能存在异常。

品牌零售价格：

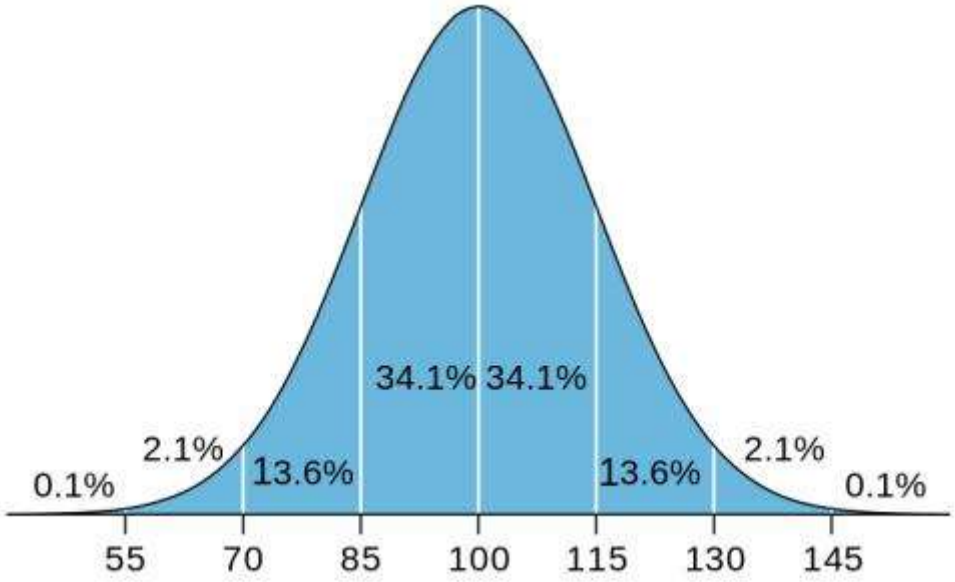


数据中共有 150 个品牌，按照某一品牌平均价格排序得到上图，价格表见下：

前十最低价格品牌		前十最高价格品牌	
品牌	**价格**	**品牌**	**价格**
广东凌鹰	239	德赛视听	2980
中兴	260	GFIVE	2980
EXUN	280	华旗爱国者	3000
大显	318	迈峰	3222.857
Tfnet	357.666667	海尔	3357.333
易丰展业	359.333333	本为	3780
康佳	366.333333	深圳科盛	3880
振华	366.5	kyocera	3980
万利达	373	35Phone	4880
TCL 移动	379	Jadeway	5155

数据离散化

为了之后对数据挖掘，需要预先对数据离散化。
根据数据量，选择离散化的数据有芯片主频，零售价格，厚度，屏幕数量，产品重量，屏幕尺寸，分辩率，RAM，ROM，Flash 内存，摄像头，电池容量。
去除异常点：假设连续数据分布符合高斯分布，根据高斯分布面积将连续数据高于或低于 3 个标准差的数据删去。



使用 python 处理得到：

```
def delect(df):
    df_copy=df
    types = '芯片主频，零售价格，厚度，屏幕数量，产品重量，屏幕尺寸，分辩率，RAM，ROM，Flash 内存，摄像头，电池容量'
    types = types.split('，')
    for type in types:
        down=df[type].mean()-3*df[type].std()
        up=df[type].mean()+3*df[type].std()
        df=df.query(f'{type}<{up} & {type}>{down}')
        print(type,df.iloc[:,0].size)
        return df
```

得到 1127 条数据。
对这 1127 条数据实现的离散：

等宽离散