目录

01 **数据探索**

- 数据质量分析
- 数据特征分析

02 **数据预处理**

- 数据抽样
- 数据清洗
- 数据变换

资料整理
火龙果软件
荐

资料整理
火龙果软件
荐

目录

数据探索

- 数据质量分析
- 数据特征分析

数据预处理

- 数据抽样
- 数据清洗
- 数据变换

01

数据探索

数据质量分析

数据质量分析

资料整理
火龙果软件
荐

数据质量分析包括很多内容，这里我们主要介绍缺失值分析和异常值分析

缺失值分析内容

1）缺失值的属性有哪些

2）属性的缺失数

3）缺失率

异常值分析方法

1）简单统计量分析

2）三倍标准差原则

3）箱型图分析

缺失值分析

资料整理
火龙果软件
荐

titanic.csv是数据挖掘的典型案例，对其进行缺失值分析

	A	B
1	=file("D:/KDD/titanic.csv").import@qtc()	/导入xls数据
2	=A1.fname()	/数据的属性
3	=A2.((y=~,A1.align@a([true,false],!eval(y))))	/按照是否缺失分组
3	=A2.new(~:col,A3(#)(1).len():null_no,A3(#)(2).len():no_null,round(null_no/A1.len(),3):null_rate)	
4	/统计属性的缺失数，未缺失数，缺失率。	

A1

	ssengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
4	4	1	1	Futrelle, Mr...	female	35	1	0	113803	53.1	C123	S
5	5	0	3	Allen, Mr. ...	male	35	0	0	373450	8.05	(null)	S
6	6	0	3	Moran, Mr. ...	male	(null)	0	0	330877	8.4583	(null)	Q

A2

	Member
1	assengerId
2	Survived
3	Pclass
4	Name
5	Sex
6	Age

A3

Index	Member
4	[[],[[1,0,3, ...],[2,1,1, ...],[3,1,3, ...], ...]]
5	[[],[[1,0,3, ...],[2,1,1, ...],[3,1,3, ...], ...]]
6	[[[6,0,3, ...],[18,1,2, ...],[20,1,3, ...], ...],[[1,0,3, ...],[2,

Index	Member
1	[[6,0,3, ...],[18,1,2, ...],[20,1,3, ...], ...]
2	[[1,0,3, ...],[2,1,1, ...],[3,1,3, ...], ...]

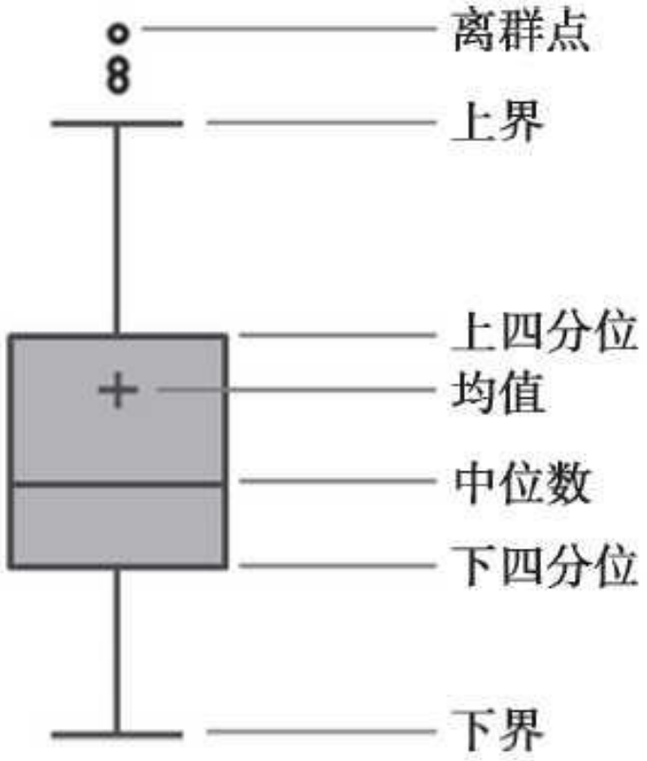
Index	PassengerId	Survived	Pclass	Name	Sex	Age	S
1	6	0	3	Moran, Mr. ...	male	(null)	
2	18	1	2	Williams, ...	male	(null)	
3	20	1	3	Masselma...	female	(null)	

A4

	col	null_no	no_null	null_rate
1	PassengerId	0	891	0.0
2	Survived	0	891	0.0
3	Pclass	0	891	0.0
4	Name	0	891	0.0
5	Sex	0	891	0.0
6	Age	177	714	0.199
7	SibSp	0	891	0.0
8	Parch	0	891	0.0
9	Ticket	0	891	0.0
10	Fare	0	891	0.0
11	Cabin	687	204	0.771
12	Embarked	2	889	0.002

异常值分析

箱形图



使用箱型图原理找到catering_sale.csv中销量的异常值

	A	B
1	=file("D:/KDD/catering_sale.csv").import@tc()	
2	=A1.(sales).median(:4)	/返回数据分4份的各分位点
3	=A2(3)-A2(1)	/四分位距
4	=A2(1)-1.5*A3	/下四分位数
5	=A2(3)+1.5*A3	/上四分位数
6	=A1.select(sales<A4||sales>A5)	/选出异常值

A1~A6结果

Index	date	sales
1	2015-03-01	51
2	2015-02-28	2618.2
3	2015-02-27	2608.4
4	2015-02-26	2651.9

Index	Member
1	2450.1
2	2655.8
3	3023.8

Value
573.7000000000003

Value
1589.5499999999995

Value
3884.3500000000004

Index	date	sales
1	2015-03-01	51
2	2015-02-21	6607.4
3	2015-02-20	4060.3
4	2015-02-14	(null)
5	2015-02-12	865
6	2014-11-08	22
7	2014-11-01	60
8	2014-09-27	9106.44
9	2014-09-08	4065.2

资料整理
火龙果软件
荐

目录

数据探索

- 数据质量分析
- 数据特征分析

数据预处理

- 数据抽样
- 数据清洗
- 数据变换

01

数据探索

数据特征分析

数据特征分析

资料整理
火龙果软件
荐

数据特征分析的内容：

- 分布分析：揭示数据的分布特征和分布类型
- 对比分析：把两个相互联系的指标进行比较
- 统计量分析：用统计指标对定量数据进行统计描述
- 相关性分析：分析连续变量之间相关程度的强弱
- ……

分布分析 - 定量数据

对于定量数据的分布分析，通常是通过绘制频率分布图观察。

house_prices.csv是房价预测的数据，请对房价（SalePrice）做出相应的分析。

先列出频率分布表，再绘制频率分布图

	A	B	C
1	=file("D:/KDD/house_prices.csv").import@qtc(Id,SalePrice)	30	/导入数据,确定组数30
2	=A1.(SalePrice)		/房价序列
3	=ceil((A2.max()-A2.min())/B1)		/计算组距
4	=B1.([(~-1)*A3+A2.min(),~*A3+A2.min()])		/等距分组
5	=A4.new(~:group,(~(1)+~(2))/2:group_median, count(A2.(~>=group(1)&&~<group(2))):count,count/A1.len():fre,cum(fre):cum_fre)		/计算频数，频率，累计频率

A1

Index	Id	SalePrice
1	1	208500
2	2	181500
3	3	223500
4	4	140000

A2

Index	Member
1	208500
2	181500
3	223500
4	140000

A3

Value
14402.0

A4

Index	Member
1	[34900.0,49302.0]
2	[49302.0,63704.0]
3	[63704.0,78106.0]
4	[78106.0,92508.0]

A5

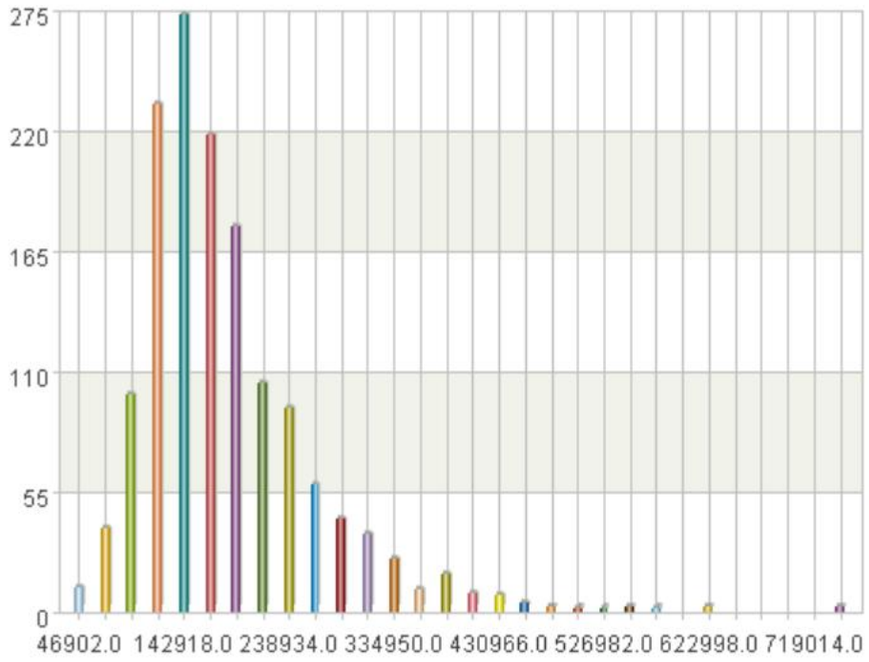
Index	group	group_me...	count	fre	cum_fre
1	[34900.0,49302.0]	42101.0	5	0.0034246...	0.0034246...
2	[49302.0,63704.0]	56503.0	11	0.0075342...	0.0109589...
3	[63704.0,78106.0]	70905.0	13	0.0089041...	0.0198630...
4	[78106.0,92508.0]	85307.0	61	0.0417808...	0.0616438...

分布分析 - 定量数据

资料整理 火龙果软件 荐

根据分组后的频率分布表，绘制频率分布直方图

	A
1	=call("E:/esproc_test/KDD/quantify_distr.dfx")
2	=canvas()
3	=A2.plot("BackGround")
4	=A2.plot("EnumAxis","name":"x")
5	=A2.plot("NumericAxis","name":"y","location":2)
6	=A2.plot("Column","axis1":"x","data1":A1.(string(group_median)),"axis2":"y","data2":A1.(count))
7	=A2.draw@p(600,450)
	A1:调用等距分组的分析程序得到分析结果；A2:定义画布；A3:定义背景；A4:绘制横轴；A5:绘制纵轴 A6:绘制柱图元；A7:画图

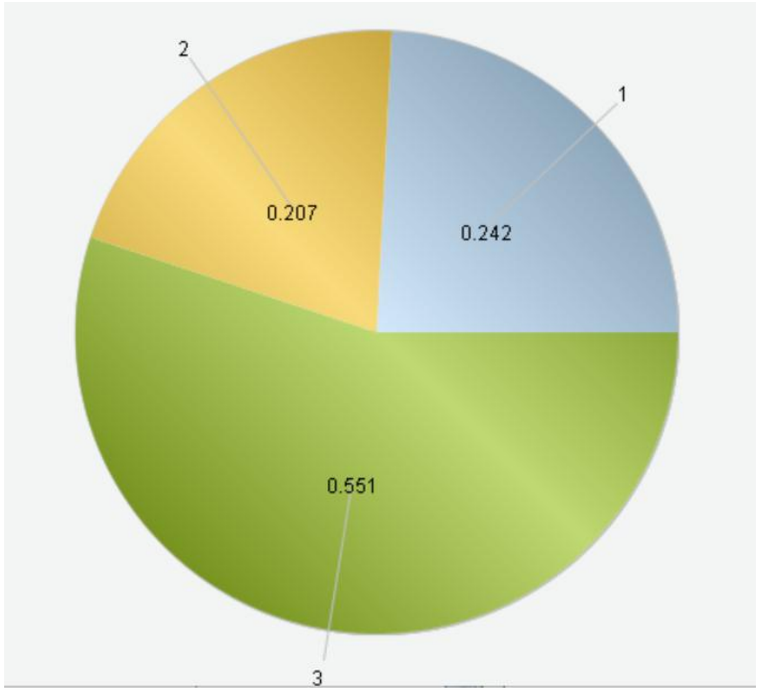


观察频率分布直方图，可以看到其呈右偏态，在预处理时需要对其进行纠偏处理。

分布分析 - 定性数据

对于定性变量，可以采用饼图来描述其分布。请绘制饼图分析titanic数据中船票等级Pclass频数分布情况。

	A
1	=file("D:/KDD/titanic.csv").import@qtc()
2	=A1.groups(Pclass;round(count(~)/A1.len(),3):Rate)
3	=canvas()
4	=A3.plot("EnumAxis","name":"x","location":3,"polarX":0.55, "allowLabels":false)
5	=A3.plot("NumericAxis","name":"y","location":4,"allowLabels":false)
6	=A3.plot("Sector","axis1":"x","data1":A2.(string(Pclass)),"axis2":"y", "data2":A2.(Rate),"text":A2.(Rate))
7	=A3.draw@p(600,450)
	/A1:导入数据；A2:计算各等级所占比例；A3:定义画布；A4:定义背景；A5:绘制极轴；A6:绘制角轴；A7:绘制环图元；A8:画图



A1

	ssengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
	1	0	3	Braund, Mr....	male	22	1	0	A/5 21171	7.25	(null)	S
2	2	1	1	Cumings, ...	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, ...	female	26	0	0	STON/O2. ...	7.925	(null)	S

A2

	Pclass	Rate
	1	0.242
2	2	0.207
3	3	0.551

对比分析

depart_sales.csv是三个部门的销售额数据，请使用折线图对比甜品部、海鲜部和素菜部2018年的销售情况。

A1

	date	dessertA	seafoodB	vegetableC
1	2016-01-01	77876	52076	46449
2	2016-01-02	81542	53742	46920
3	2016-01-03	75441	55259	52641

	A	B
1	=file("D:/KDD/depart_sales.csv").import@qtc()	/导入数据
2	=A1.select(year(date)==2018)	/筛选2018年数据
3	=A2.groups(month(date):date;avg(dessertA):dessertA,avg(seafoodB):seafoodB,avg(vegetableC):vegetableC)	/按月份组求平均
4	=canvas()	/定义画布
5	=A4.plot("EnumAxis","name":"x","allowRegions":false)	/绘制横轴
6	=A4.plot("NumericAxis","name":"y","location":2,"autoCalcValueRange":true)	/绘制纵轴
7	=A4.plot("Line","lineColor":-32768,"markerColor":["ChartColor",0,true,-8355840,-14336,0],"axis1":"x","data1":A3.("2018-"+string(date)),"axis2":"y","data2":A3.(dessertA))	/绘制线图元
8	=A4.plot("Line","lineColor":-8388480,"markerColor":["ChartColor",0,true,-8355840,-65281,0],"axis1":"x","data1":A3.("2018-"+string(date)),"axis2":"y","data2":A3.(seafoodB))	/绘制线图元
9	=A4.plot("Line","lineColor":-6724045,"markerColor":["ChartColor",0,true,-8355840,-16776961,0],"axis1":"x","data1":A3.("2018-"+string(date)),"axis2":"y","data2":A3.(vegetableC))	/绘制线图元
10	=A4.plot("Text","text":"dessetA","x":-80,"y":65,"textStyle":1,"textColor":-32768)	/绘制文本图元
11	=A4.plot("Text","text":"seafoodB","x":-80,"y":30,"textStyle":1,"textColor":-8388480)	/绘制文本图元
12	=A4.plot("Text","text":"vegetableC","x":-80,"y":100,"textStyle":1,"textColor":-16777216)	/绘制文本图元
13	=A4.draw@p(450,250)	/画图

对比分析

绘制折线图比较海鲜部三年内的销量变化。

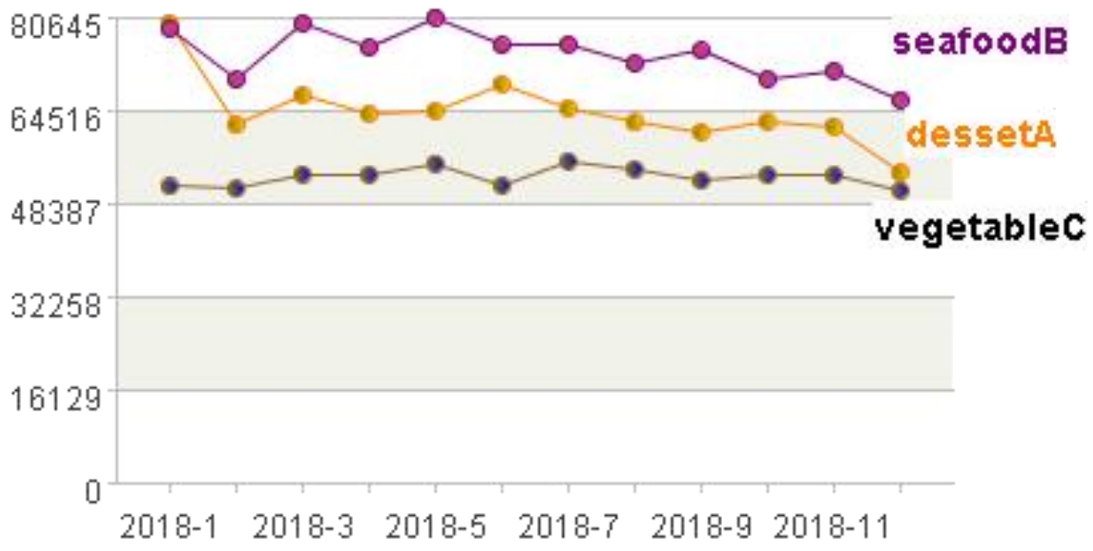
A1

	date	dessertA	seafoodB	vegetableC
1	2016-01-01	77876	52076	46449
2	2016-01-02	81542	53742	46920
3	2016-01-03	75441	55259	52641

	A	B
1	=file("D:/KDD/depart_sales.csv").import@qtc()	/导入date，seafoodB
2	=A1.group(year(date))	/按年分组
3	=A2.(~.groups(year(date):year,month(date):month;avg(seafoodB):seafoodB))	/自己按月分组，求平均
4	=canvas()	/定义画布
5	=A4.plot("EnumAxis","name":"x","allowRegions":false)	/绘制横轴
6	=A4.plot("NumericAxis","name":"y","location":2,"autoCalcValueRange":true)	/绘制纵轴
7	=A4.plot("Line","lineColor":-32768,"markerColor":["ChartColor",0,true,-8355840,-14336,0],"axis1":"x","data1":A3(1).(string(month)),"axis2":"y","data2":A3(1).(seafoodB))	/绘制线图元
8	=A4.plot("Line","lineColor":-8388480,"markerColor":["ChartColor",0,true,-8355840,-65281,0],"axis1":"x","data1":A3(2).(string(month)),"axis2":"y","data2":A3(2).(seafoodB))	/绘制线图元
9	=A4.plot("Line","lineColor":-6724045,"markerColor":["ChartColor",0,true,-8355840,-16776961,0],"axis1":"x","data1":A3(3).(string(month)),"axis2":"y","data2":A3(3).(seafoodB))	/绘制线图元
10	=A4.plot("Text","text":"2016","x":-80,"y":80,"textStyle":1,"textColor":-32768)	/绘制文本图元
11	=A4.plot("Text","text":"2017","x":-80,"y":65,"textStyle":1,"textColor":-8388480)	/绘制文本图元
12	=A4.plot("Text","text":"2018","x":-80,"y":40,"textStyle":1,"textColor":-16777216)	/绘制文本图元
13	=A4.draw@p(450,250)	/画图

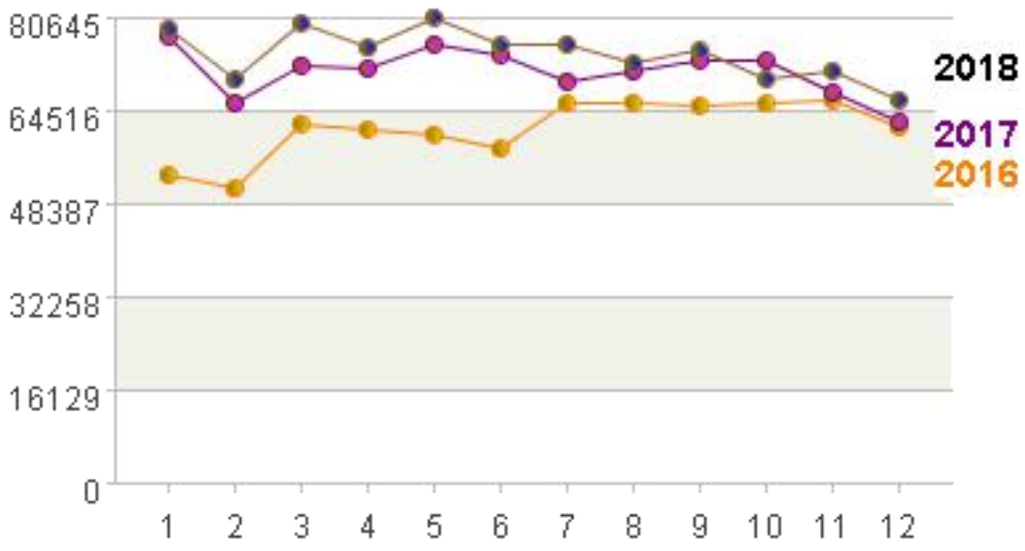
资料整理
火龙果软件
荐

对比分析



2018年三部门对比图

总体看甜品部A和海鲜部B销量呈下降趋势，而素材部相较于甜品部A和海鲜部B变化平稳。进而可以进一步分析其中原因。。



海鲜部三年对比图

总体看海鲜部B三年的销量是逐年上升的，但17年和18年又在本年度呈现下降趋势，可以从其他的角度分析其中原因。

对比分析

绘制饼图比较性别是否影响幸存。

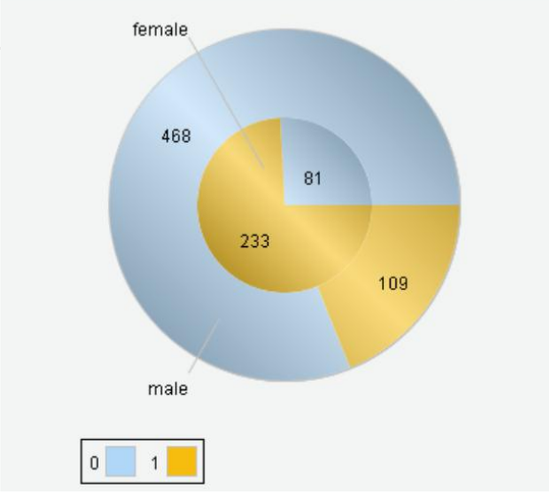
	A	B
1	=file("D:/KDD/titanic.csv").import@qtc()	/导入数据
2	=A1.groups(Sex,Survived;count(~):Cnt)	/按照性别和是否幸存分组计数
3	=canvas()	/定义画布
4	=A3.plot("EnumAxis","name":"x","location":3,"polarX":0.55,"polarY":0.4,"allowLabels":false)	/绘制极轴
5	=A3.plot("NumericAxis","name":"y","location":4,"allowLabels":false)	/绘制角轴
6	=A3.plot("Sector","axis1":"x","data1":A2.(Sex+","+string(Survived)),"axis2":"y","data2":A2.(Cnt),"text":A2.(Cnt))	/绘制环图元
7	=A3.plot("Legend","legendText":A1.id(string(Survived)),"columns":2,"x":0.2,"y":0.8)	/绘制图例
8	=A3.draw@p(400,400)	/画图

A1

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr....	male	22	1	0	A/5 21171	7.25	(null)	S
2	2	1	1	Cumings, ...	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, ...	female	26	0	0	STON/O...	7.925	(null)	S

A2

	Sex	Survived	Cnt
1	female	0	81
2	female	1	233
3	male	0	468
4	male	1	109



观察饼图发现，男性幸存者所占比例远远小于女性幸存者，说明性别与是否幸存关系密切，应该把它当做重要变量。

对比分析

绘制柱状图比较船票等级是否影响幸存。

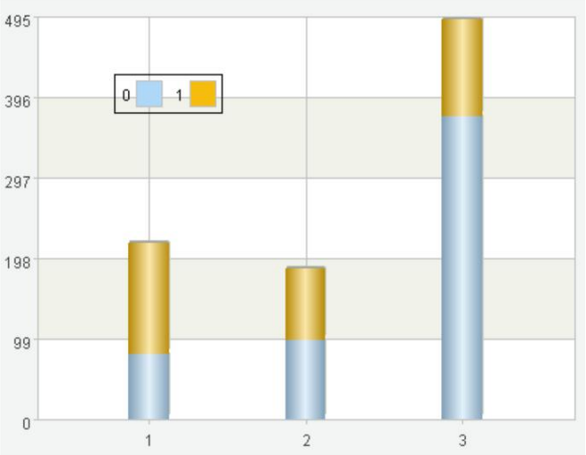
	A	B
1	=file("D:/KDD/titanic.csv").import@qtc()	/导入数据
2	=A1.groups(Pclass,Survived;count(~):Cnt)	/按照船票等级和是否幸存分组计数
3	=canvas()	/定义画布
4	=A3.plot("EnumAxis","name":"x")	/绘制横轴
5	=A3.plot("NumericAxis","name":"y","location":2)	/绘制纵轴
6	=A3.plot("Column","stackType":2,"axis1":"x","data1":A2.(string(Pclass)+","+string(Survived)),"axis2":"y","data2":A2.(Cnt))	/绘制柱图元
7	=A3.plot("Legend","legendText":A1.id(string(Survived)),"columns":2,"x":0.2)	/绘制图例
8	=A3.draw@p(600,450)	/画图

A1

Index	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr....	male	22	1	0	A/5 21171	7.25	(null)	S
2	2	1	1	Cumings, ...	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, ...	female	26	0	0	STON/O...	7.925	(null)	S

A2

Index	Pclass	Survived	Cnt
1	1	0	80
2	1	1	136
3	2	0	97
4	2	1	87
5	3	0	372
6	3	1	119



观察柱状图发现，人数最多的是第3等级的，但幸存者比例是最低的，而第一等级的幸存比例是最高的，说明船票等级也会影响是否幸存。

统计量分析

对于定量数据，通常需要对其进行统计量分析，请计算房价数据的统计量。

	A	B
1	=file("D:/KDD/house_prices.csv").import@qtc(SalePrice)	/导入数据
2	=A1.(SalePrice)	/取出房价序列
3	=A2.max()	/最大值
4	=A2.min()	/最小值
5	=A2.avg()	/平均值
6	=A2.median(:4)	/四等分后的分界点
7	=power(A2.sum(power(~-A5,2))/(A2.len()-1),0.5)	/标准差（样本标准差，除n-1）
8	=A2.sum(power(~-A5,3))*A2.len()/((A2.len()-1)*(A2.len()-2)*power(A7,3))	/偏度

Value
755000

Value
34900

Value
180921.19589...

Index	Member
1	129950.0
2	163000.0
3	214000.0

Value
79442.50288288663

Value
1.88287575976821

A1

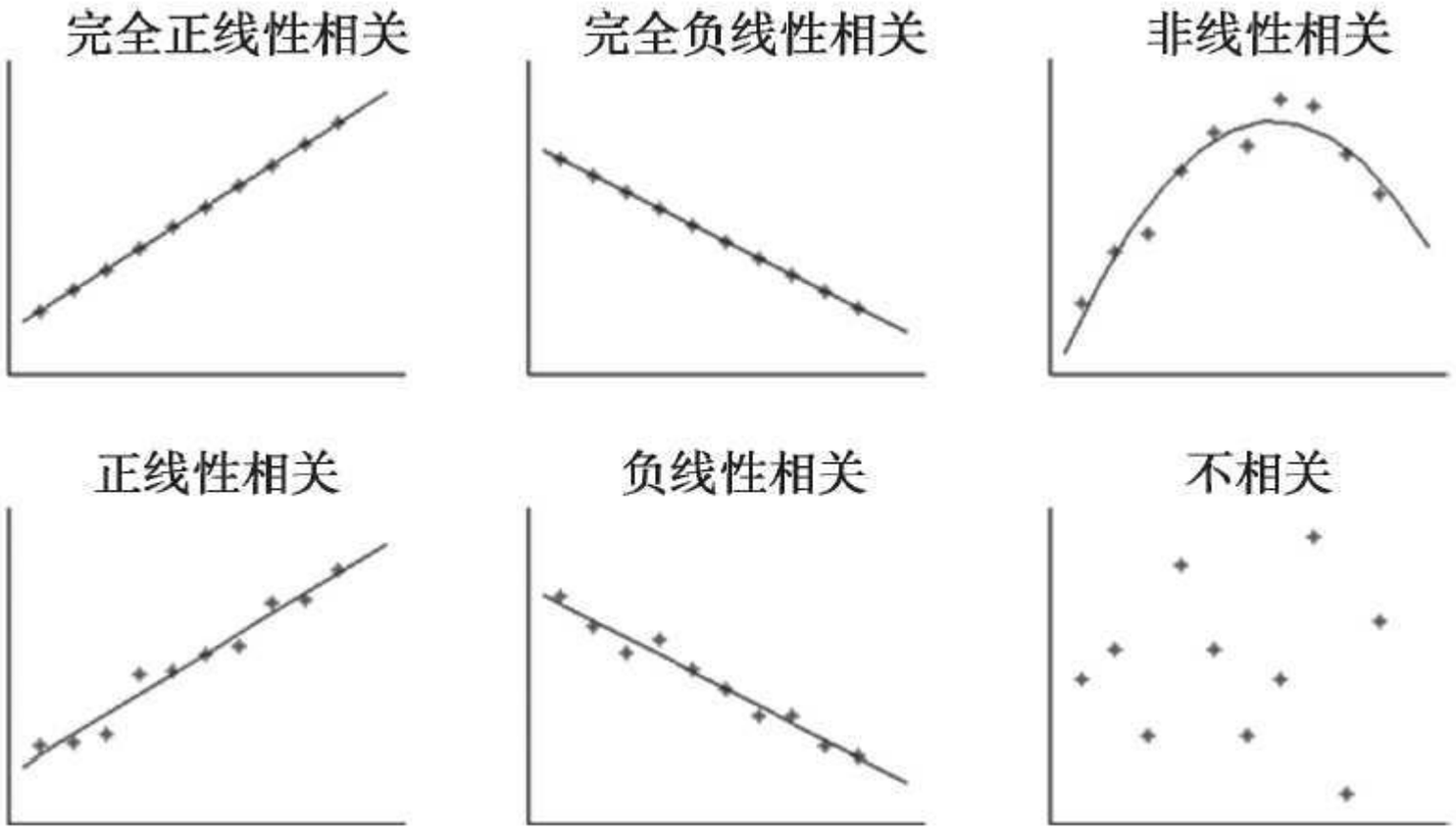
Index	Id	SalePrice
1	1	208500
2	2	181500
3	3	223500
4	4	140000

A2

Index	Member
1	208500
2	181500
3	223500
4	140000

相关性分析 - 两个定量数据

判断两个变量是否具有线性相关关系的最直观的方法是直接绘制散点图。



相关性分析 - 两个定量数据

请绘制散点图说明生活面积和房价之间的关系。

	A	B
1	=file("D:/KDD/house_prices.csv").import@qtc(GrLivArea,SalePrice)	/导入数据
2	=A1.(#1)	/生活面积序列
3	=A1.(#2)	/房价序列
4	=(A2.len()*A2.sum(~*A3(#))-A2.sum()*A3.sum())/(A2.len()*A2.sum(~*~)-power(A2.sum(),2))	/最小二乘法计算斜率
5	=A3.avg()-A4*A2.avg()	/计算截距
6	=[A2.min(),A2.max()]	/取生活面积两个极值
7	=A6.(A4*~+A5)	/计算直线的两个极值
8	=canvas()	/定义画布
9	=A8.plot("NumericAxis","name":"x","title":"SalePrice")	/绘制横轴
10	=A8.plot("NumericAxis","name":"y","location":2,"title":"GrLivArea")	/绘制纵轴
11	=A8.plot("Dot","lineWeight":0,"lineColor":-16776961,"markerWeight":1,"axis1":"x","data1":A2,"axis2":"y","data2": A3)	/绘制点图元
12	=A8.plot("Line","markerStyle":0,"lineColor":-65536,"axis1":"x","data1":A6,"axis2":"y","data2":A7)	/绘制线图元
13	=A8.draw@p(600,450)	/画图

A1

Index	GrLivArea	SalePrice
1	1710	208500
2	1262	181500
3	1786	223500
4	1717	140000

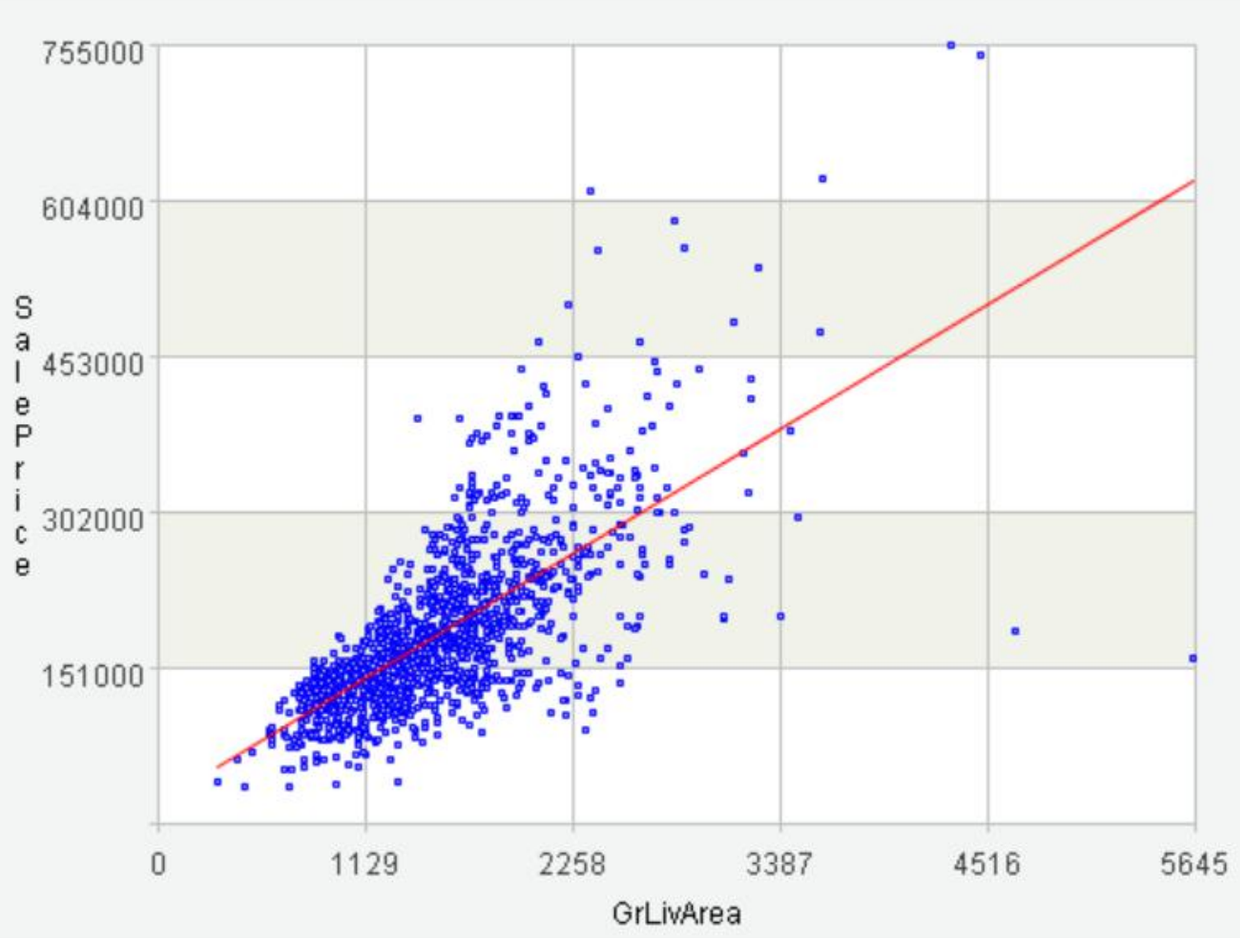
A4

Value
107.13035896582518

A5

Value
18569.02585648725

相关性分析 - 两个定量数据



观察散点图，总体上呈现生活面积越大房价越高的趋势，说明两者相关性很高。但右下角两个点很特殊，生活面积很大，但房价确很低，影响了整体的线性关系，因此可以将其视为异常点删除。

相关性分析 - 两个定量数据

资料整理
火龙果软件
荐

Pearson相关系数：

$$r(x_1,x_2)=\frac{\sum_{i=1}^{n}\left(x_1(i)-\bar{x}_1\right)\left(x_2(i)-\bar{x}_2\right)}{\left(\sum_{i=1}^{n}\left(x_1(i)-\bar{x}_1\right)^2\sum_{i=1}^{n}\left(x_2(i)-\bar{x}_2\right)^2\right)^{1/2}}$$

Spearman秩相关系数：

$$\theta=\frac{\sum_{i=1}^{n}\left(R_i-\bar{R}\right)\left(S_i-\bar{S}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_i-\bar{R}\right)^2\sum_{i=1}^{n}\left(S_i-\bar{S}\right)^2}}$$

只要两个变量具有严格单调的函数关系，那么它们就是完全Spearman相关的，Pearson相关只有在变量具有线性关系时才是完全相关的。

相关性分析 - 两个定量数据

计算生活面积和房价之间的相关系数。

A1

Index	GrLivArea	SalePrice
1	1710	208500
2	1262	181500
3	1786	223500
4	1717	140000

	A	B	C	
1	=file("D:/KDD/house_prices.csv").import@qtc(GrLivArea,SalePrice)			
2	=A1.(#1)			
3	=A1.(#2)			
4	=func(A6,A2,A3)	Value 0.7086244776126511		
5	=func(A6,A2,A3,"spearman")		Value 0.7313095834659141	
6	func			
7		if C6=="pearson"||C6==null	=A6.avg()	
8			=B6.avg()	
9			=A6.sum((~-C7)*(B6(#)-C8))	
10			=power(A6.sum(power(~-C7,2))*B6.sum(power(~-C8,2)),0.5)	
11			return C9/C10	
12		else if C6=="spearman"	=A6.ranks@s()	/计算排名
13			=B6.ranks@s()	/计算排名
14			return func(A6,C12,C13,"pearson")	

A1:读入生活面积和房价数据
A2:生活面积序列
A3:房价序列
A4:计算pearson相关系数
A5:计算spearman相关系数
A6:按公式计算两个序列的pearson或者spearman

两个相关系数都大于0.7，说明两者的线性关系很强，和散点图分析的结果一致。

相关性分析 - 两个定性数据

	C1	C2	...	Cj	合计
R1	f11	f12	...	f1j	fr1
R2	f21	f22	...	f2j	fr2
...	...	...	...	...	
Ri	fi1	fi2	...	fij	fri
合计	fc1	fc2	...	fcj	sum

计算卡方值:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o ： 实际值

$f_e = f_{ri} * f_{cj}/sum$： 期望值

计算V相关系数:

$$V = \sqrt{\frac{\chi^2}{n * min([R-1, C-1])}}$$

R： 定性数据1的类别数

C： 定性数据2的类别数

相关性分析 - 两个定性数据

计算性别和是否幸存之间的V相关系数。

	A	B	C
1	=file("D:/KDD/titanic.csv").import@qtc(Sex,Survived)		/导入数据
2	=func(A3,A1,"Sex","Survived")		/计算V相关系数
3	func		
4		=A3.groups(${B3},${C3};count(~):cnt)	/按照两个变量分组计数
5		=B4.group(${B3}).(~.sum(cnt))	/按第一个变量分组计数
6		=B4.group(${C3}).(~.sum(cnt))	/按第二个变量分组计数
7		=B5.conj(B6.(~*B5.~/B5.sum()))	/计算期望值
8		=power(B4.(cnt).sum(power(~-B7(#),2)/B7(#))/(B5.sum()*min([B5.len()-1,B6.len()-1])),0.5)	
		/根据公式计算V相关系数	

Value
0.5433513806577551

Index	Sex	Survived
1	male	0
2	female	1
3	female	1
4	female	1
5	male	0

性别与是否幸存之间的相关系数是0.54，和我们之前分析的结果一样，性别对于是否幸存很重要。

资料整理
火龙果软件
荐

目录

数据探索

- 数据质量分析
- 数据特征分析

数据预处理

- 数据抽样
- 数据清洗
- 数据变换

02

数据预处理

数据抽样

数据取样

资料整理
火龙果软件
荐

数据量特别大时，对数据进行抽样，可以更好的提取数据的相关信息，也可以提高建模效率。

抽样数据的方法

- 随机抽样
- 等距抽样
- 分层抽样
- 起始顺序抽样
- 分类抽样

使用不同的抽样方式抽取样本

A7

Index	Member
1	[[4,1,1, ...],[8,4,1, ...],[14,6,1, ...], ...]
2	[[1,6,1, ...],[2,7,0, ...],[3,6,1, ...], ...]

Index	ID	DayOfWeek	Open	Promo	Holiday	Sales
72177	305...	4	1	0	0	high
72178	305...	4	1	0	1	high
72179	305...	2	1	1	0	high

A8

Index	ID	DayOfWeek	Open	Promo	Holiday	
72176	259099	6	1	0	0	low
72177	188535	5	0	0	1	low
72178	112718	1	1	0	0	low

	A	B
1	=file("D:/datamining/data_exploration/sales_c.csv").import@tc()	/导入数据
2	=A1.sort(rand()).m(:A1.len()*0.1)	/随机抽取10%，随机抽样
3	=A1.select(#%20==0)	/等距抽取5%，等距抽样
4	=A1.group(DayOfWeek)	/按照星期分层
5	=A4.(if(#<=5,~.sort(rand()).m(:~.len()*0.1),~.sort(rand()).m(:~.len()*0.2))).conj()	/周一至周五抽10%,周末抽20%,分层抽样
6	=A1.select(#<=A1.len()*0.3)	/从起始顺序抽样30%
7	=A1.group(Sales)	/按销量高低分类
8	=A7.(~.sort(rand()).m(:A7(1).len()*0.5)).conj()	/销量高抽取50%,销量低数量与销量高一样,分类抽样

A1

Index	ID	DayOfWeek	Open	Promo	Holiday	
305161	305161	2	1	1	0	high
305162	305162	2	1	0	0	low
305163	305163	2	1	1	0	low

A4

Index	Member
1	[[4,1,1, ...],[7,1,1, ...],[9,1,1, ...], ...]
2	[[5,2,1, ...],[19,2,1, ...],[29,2,1, ...], ...]

Index	ID	DayOfWeek	Open	Promo	Holiday	Sales
43223	305145	1	1	1	0	high
43224	305148	1	1	1	0	high
43225	305154	1	1	1	0	low

A2

Index	ID	DayOfWeek	Open	Promo	Holiday	
30514	107448	1	1	0	1	low
30515	98195	7	0	0	0	low
30516	273497	4	1	0	0	low

A3

Index	ID	DayOfWeek	Open	Promo	Holiday	
15256	305120	3	1	0	0	low
15257	305140	5	1	1	1	low
15258	305160	7	0	0	0	low

A5

Index	ID	DayOfWeek	Open	Promo	Holiday	
39207	221610	7	0	0	1	low
39208	186706	7	0	0	0	low
39209	85702	7	0	0	0	low

A6

Index	ID	DayOfWeek	Open	Promo	Holiday	
91546	91546	4	1	1	0	low
91547	91547	1	1	0	0	high
91548	91548	3	1	0	0	low

资料整理
火龙果软件
荐

目录

数据探索

- 数据质量分析
- 数据特征分析

数据预处理

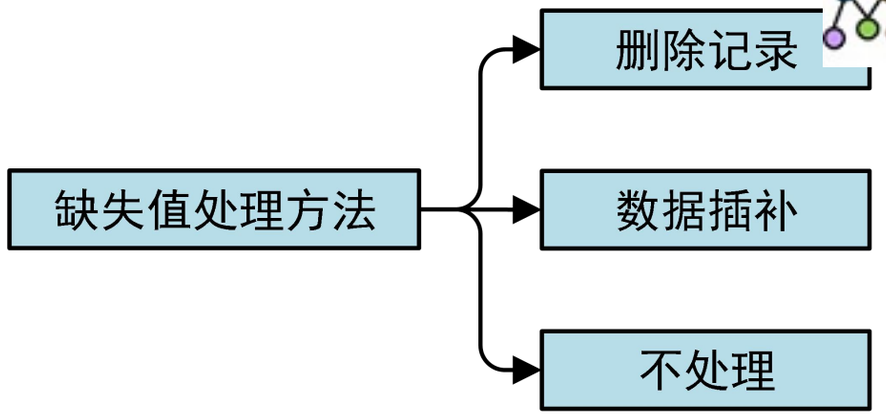
- 数据抽样
- 数据清洗
- 数据变换

02

数据预处理

数据清洗

缺失值处理



插补方法	方法描述
均值/中位数/众数插补	根据属性值的类型，用该属性值的平均数/中位数/众数进行插补
使用固定值	将缺失的属性值用一个常量替换。如用某地外来务工人员的平均工资来插补某位外来务工人员的“基本工资”。
最近邻插补	在记录中找到与缺失样本最接近的的样本的该属性值插补
回归方法	对带有缺失值的变量，根据已有数据和与其有关的其他变量的数据建立拟合模型来预测缺失值的属性值
插值法	插值法是利用已知点建立合适的插值函数f(x)，未知值由对应点xi求出的函数值f(xi)近似代替

缺失值填补——简单方式

	A	均值插补
1	=file("D:/KDD/titanic.csv").import@qtc(PassengerId,Name,Age)	/导入数据
2	=A1.group(!Age)	/按照是否缺失分组
3	=A2(2).run(Age=A2(1).avg(Age))	/使用不缺失那组的均值填补缺失的那组

A1

	PassengerId	Name	Age
17	17	Rice, Mast...	2
18	18	Williams, ...	(null)
19	19	Vander Pla...	31
20	20	Masselma...	(null)

A2

1	[[1,Braund, Mr
2	[[6,Moran, Mr.

Index	PassengerId	Name	Age
1	6	Moran, Mr. ...	(null)
2	18	Williams, ...	(null)
3	20	Masselma...	(null)
4	27	Emir, Mr. F...	(null)

A3

	PassengerId	Name	Age
1	6	Moran, Mr. ...	29.699117...
2	18	Williams, ...	29.699117...
3	20	Masselma...	29.699117...
4	27	Emir, Mr. F...	29.699117...

填补后的A1

	Id	Name	Age
18	18	Williams, ...	29.699117...
19	19	Vander Pla...	31
20	20	Masselma...	29.699117...
21	21	Fynney, Mr....	35

	A	众数插补
1	=file("D:/KDD/titanic.csv").import@qtc(PassengerId,Embarked)	/导入数据
2	=A1.group(Embarked)	/按照是否缺失分组
3	=A2(1).run(Embarked=A2.maxp(~.len()).Embarked)	/使用众数填补缺失的那组

A1

	PassengerId	Embarked
61	61	C
62	62	(null)
63	63	S
64	64	S

A2

	Member
1	[[62,],[830,]]
2	[[2,C],[10,C],[20,C], ...]
3	[[6,Q],[17,Q],[23,Q], ...]
4	[[1,S],[3,S],[4,S], ...]

Index	PassengerId	Embarked
1	62	(null)
2	830	(null)

A3

	PassengerId	Embarked
1	62	S
2	830	S

填补后的A1

		Embarked
61	61	C
62	62	S
63	63	S
64	64	S

缺失值填补 - 根据内容填补

资料整理
火龙果软件
荐

Mr，Miss，Mrs这些称呼在一定程度上反映了年龄，因此我们可以取这些称呼的平均值对填补Age的缺失值，这样似乎比直接使用平均值更合理一些。

	A	B
1	=file("D:/KDD/titanic.csv").import@qtc(PassengerId,Name,Age)	/导入数据
2	>A1.run(Name=Name.split@b(",")(2).split(".")(1))	/将Name改成Name中的称呼
3	>A1.group(Name).((age=~.avg(Age),~.run(Age=if(!Age,age,Age))))	/按照Name分组，使用组内均值填补缺失值

A1执行时的A1

Index	PassengerId	Name	Age
17	17	Rice, Master. Eu...	2
18	18	Williams, Mr. Ch...	(null)
19	19	Vander Planke, ...	31
20	20	Masselmani, Mrs...	(null)
21	21	Fynney, Mr. Jose...	35

A2执行后的A1

Index	PassengerId	Name	Age
17	17	Master	2
18	18	Mr	(null)
19	19	Mrs	31
20	20	Mrs	(null)
21	21	Mr	35

A3执行后的A1

Index	PassengerId	Name	Age
17	17	Master	2
18	18	Mr	32.36809045...
19	19	Mrs	31
20	20	Mrs	35.89814814...
21	21	Mr	35

缺失值填补 - 根据内容填补

资料整理
火龙果软件
荐

车库的质量信息有缺失，但是发现有不少房子的车库面积为0，表示没有车库，这种情况填补缺失值时就要把缺失值视为另一种类别如“missing/unknown”。

	A	B
1	=file("D:/KDD/house_prices.csv").import@qtc(Id,GarageArea,GarageQual)	/导入数据
2	=A1.select(GarageArea==0)	/选出车库面积为0的房子
3	>A2.run(GarageQual="missing")	/用missing填补车库质量缺失值

A1

Index	Id	GarageArea	GarageQual
88	88	528	TA
89	89	0	(null)
90	90	0	(null)
91	91	420	TA

A2

Index	Id	GarageArea	GarageQual
2	49	0	(null)
3	79	0	(null)
4	89	0	(null)
5	90	0	(null)

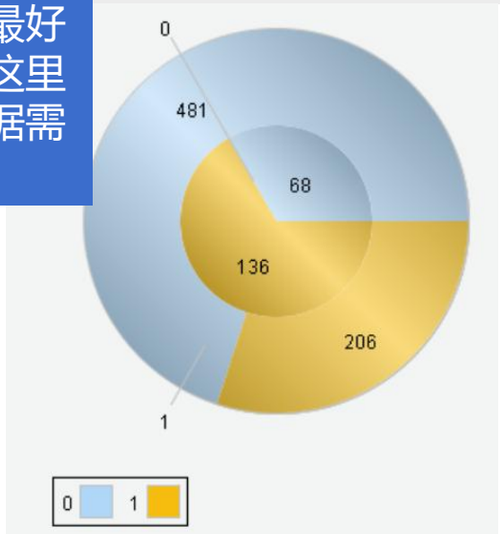
A3执行后的A1

Index	Id	GarageArea	GarageQual
88	88	528	TA
89	89	0	missing
90	90	0	missing
91	91	420	TA

缺失值处理 - 提取信息

缺失值处理没有哪个固定的方法是最好的，需要根据数据特点进行处理，这里介绍的方法都可以使用而且有些数据需要使用多种方法处理。

缺失值分析时发现Cabin缺失率高达77%，直接填补没有太大的意义，我们可以把是否缺失作为一个信息提取出来，观察其是否影响生存。



观察饼图发现缺失的生存率远低于不缺失的生存率，说明我们提取的信息是有用的。

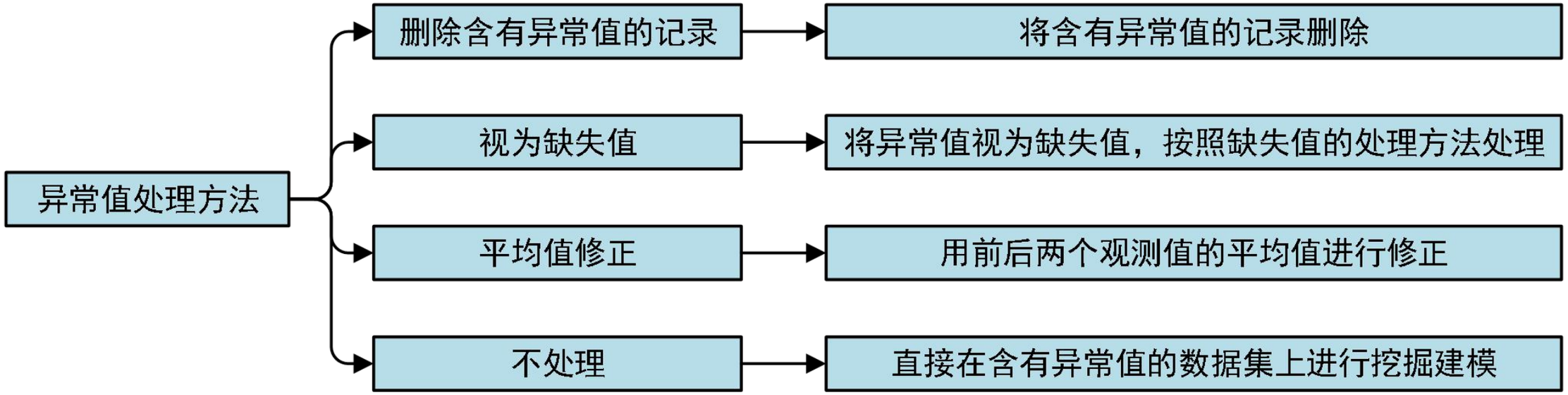
	A	B
1	=file("D:/KDD/titanic.csv").import@qtc().derive(if(!Cabin,1,0):Cabin_b)	/增加Cabin是否缺失列
2	=A1.groups(Cabin_b,Survived;count(~):Cnt)	/分组计数
3	=canvas()	/定义画布
4	=A3.plot("EnumAxis","name":"x","location":3,"polarX":0.55,"polarY": 0.4,"allowLabels":false)	/绘制极轴
5	=A3.plot("NumericAxis","name":"y","location":4,"allowLabels":false)	/绘制角轴
6	=A3.plot("Sector","axis1":"x","data1":A2.(string(Cabin_b)+","+string(Survived)),"axis2": "y","data2":A2.(Cnt),"text":A2.(Cnt))	/绘制环图元
7	=A3.plot("Legend","legendText":A1.id(string(Survived)),"columns":2,"x":0.2,"y":0.8)	/绘制图例
8	=A3.draw@p(400,400)	/画图

A1

Index	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Cabin_b
1	1	0	3	Braund, Mr....	male	22	1	0	A/5 21171	7.25	(null)	S	1
2	2	1	1	Cumings, ...	female	38	1	0	PC 17599	71.2833	C85	C	0
3	3	1	3	Heikkinen, ...	female	26	0	0	STON/O2. ...	7.925	(null)	S	1

异常值处理

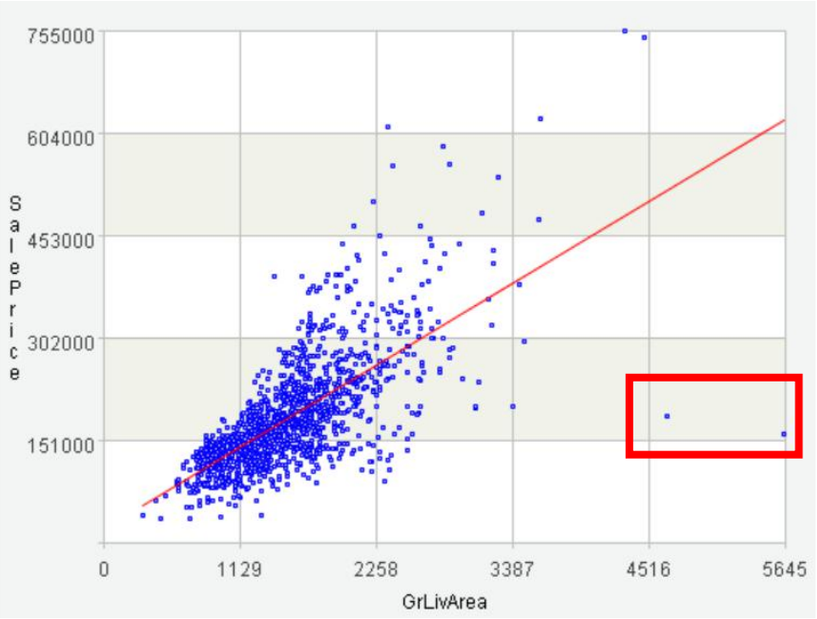
资料整理
火龙果软件
荐



在很多情况下，要先分析异常值出现的可能原因，再判断异常值是否应该舍弃，如果是正确的数据，可以直接在具有异常值的数据集上进行挖掘建模。

异常值处理 - 删除异常值

介绍数据相关性时，我们发现生活面积与房价关系密切，但有两个点很特殊，生活面积很大但房价很低，这时我们需要删除这两条记录。



	A	B
1	=file("D:/KDD/house_prices.csv").import@qtc()	/导入数据
2	=A1.top(-2,GrLivArea)(2)	/找到GrLivArea第二大的数
3	=A1.select(GrLivArea<A2)	/过滤掉两个异常值

A1

Index	Id	MSSubCla...	MSZoning	LotFrontage	LotAr
1458	1458	70	RL	66	9
1459	1459	20	RL	68	9
1460	1460	20	RL	75	9

A2

Value
4676

A3

Index	Id	MSSubCla...	MSZoning	LotFrontage	LotAr
1456	1458	70	RL	66	9
1457	1459	20	RL	68	9
1458	1460	20	RL	75	9

异常值处理和缺失值类似，没有哪种固定的方法是最好的，都需要根据数据特点进行处理。

资料整理
火龙果软件
荐

目录

数据探索

- 数据质量分析
- 数据特征分析

数据预处理

- 数据抽样
- 数据清洗
- 数据变换

02

数据预处理

数据变换

数据变换

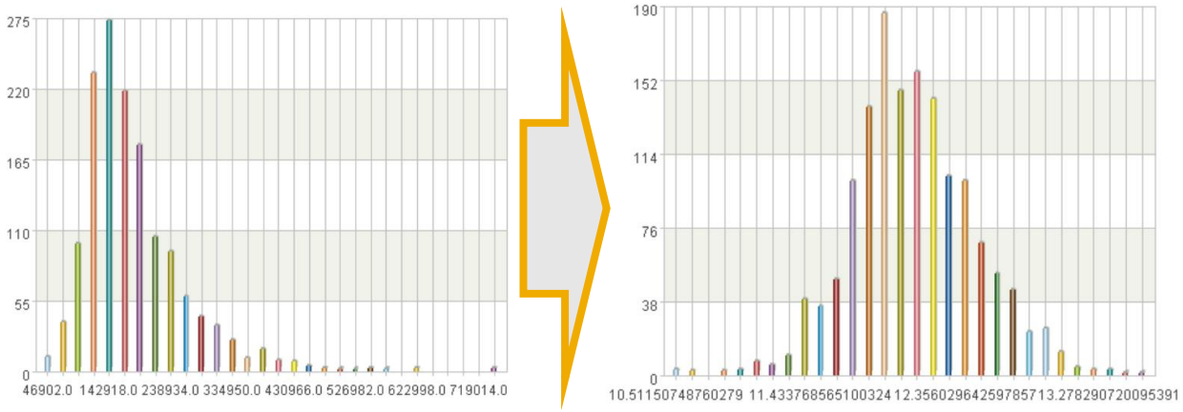
数据变换主要是对数据进行规范化处理，将数据转换成“适当的”形式，以适用于挖掘任务及算法的需要。

数据变换

- 简单函数变换
  - $x' = x^2$
  - $x' = \sqrt{x}$
  - $x' = \log(x)$
  - $\nabla f(x_k) = f(x_{k+1}) - f(x_k)$
- 规范化
  - 最小-最大规范化
  - 零-均值规范化
  - 小数定标规范化
- 连续属性离散化
  - 等宽法
  - 等频法
  - 基于聚类分析的方法
- 属性构造

对数变换进行纠偏

对房价数据进行纠偏



	A	B	C
1	=file("D:/KDD/house_prices.csv").import@qtc()	30	/导入数据
2	=A1.run(SalePrice=ln(SalePrice+1)).(SalePrice)		/纠偏并取出纠偏后的SalePrice
3	=(A2.max()-A2.min())/B1		/计算频率分布表和绘制频率分布图
4	=B1.([(~-1)*A3+A2.min(),~*A3+A2.min()])		
5	=A4.new(~:group,(~(1)+~(2))/2:group_median, count(A2.(~>=group(1)&&~<group(2))):cnt)		
6	=canvas()		
7	=A6.plot("BackGround")		
8	=A6.plot("EnumAxis","name":"x")		
9	=A6.plot("NumericAxis","name":"y","location":2)		
10	=A6.plot("Column","axis1":"x","data1":A5.(string(group_median)),"axis2":"y","data2":A5.(cnt))		
11	=A6.draw@p(600,450)		

右偏的数据经过对数变换后，数据类似正态分布了，更适合进行数据挖掘。

连续属性规范化

对房价数据进行规范化

	A	B
1	=file("D:/KDD/house_prices.csv").import@qtc(Id,SalePrice)	/导入数据
2	>A1.run(SalePrice=ln(SalePrice+1))	/对数变换纠偏
3	>A1.run(SalePrice=func(A4,A1.(SalePrice))(#))	/零——均值标准化，将数据转换成标准正态分布
4	func	=A4.avg()
5		=power(A4.sum(power(~-B4,2))/(A4.len()-1),0.5)
6		return A4.((~-B4)/B5)

A1执行时的A1

Index	Id	SalePrice
1	1	208500
2	2	181500
3	3	223500
4	4	140000
5	5	250000

A2执行后的A1

Index	Id	SalePrice
1	1	12.2476991...
2	2	12.1090164...
3	3	12.3171711...
4	4	11.8494048...
5	5	12.4292201...

A3执行后的A1

Index	Id	SalePrice
1	1	0.5598751490871431
2	2	0.186099421590100...
3	3	0.5262656270645685
4	4	-0.22788261573206...
5	5	0.5967121186204216

连续属性离散化

对地块面积LotArea进行等频离散化

	A	B	C	D
1	=file("D:/KDD/house_prices.csv").import@qtc(Id,LotArea)			/导入数据
2	=A1.(LotArea)			/提取销售额数据
3	=func(A4,A2,5,"f")			/w等宽，f等频，离散化
4	func			
5		if C4=="f"	=A4.median(:B4)|A4.max()	/C5、C6计算分割点
6		if C4=="w"	=C5=A4.iterate(~~|[~~.m(-1)+(A4.max()-A4.min())/(B4)],[A4.min()],~~.m(-1)>A4.max()).m(2:)	
7		=C5.(avg(if(!~[-1],1,~[-1]),~))		/分界点的均值作为该区间的值
8		=A4.run(~=B7(C5.pseg(~)+1))		/属于哪个区间就用区间均值替换

离散化前

Index	Member
1	8450
2	9600
3	11250
4	9550
5	14260
6	14115
7	10084
8	10382

等频离散化

Index	Member
1	7933.0
2	9495.75
3	11202.75
4	9495.75
5	113726.0
6	113726.0
7	9495.75
8	11202.75

变量交互、生成衍生变量

用SibSp+Parch生成Family

	A	B
1	=file("D:/KDD/titanic.csv").import@qtc(PassengerId,SibSp,Parch)	/导入数据
2	=A1.derive(SibSp+Parch:Family)	/生成Familylie

A1、A2

Index	PassengerId	SibSp	Parch
1	1	1	0
2	2	1	0
3	3	0	0
4	4	1	0
5	5	0	0
6	6	0	0
7	7	0	0
8	8	3	1
9	9	0	2
10	10	1	0

Index	PassengerId	SibSp	Parch	Family
1	1	1	0	1
2	2	1	0	1
3	3	0	0	0
4	4	1	0	1
5	5	0	0	0
6	6	0	0	0
7	7	0	0	0
8	8	3	1	4
9	9	0	2	2
10	10	1	0	1

变量交互、生成衍生变量

资料整理
火龙果软件
荐

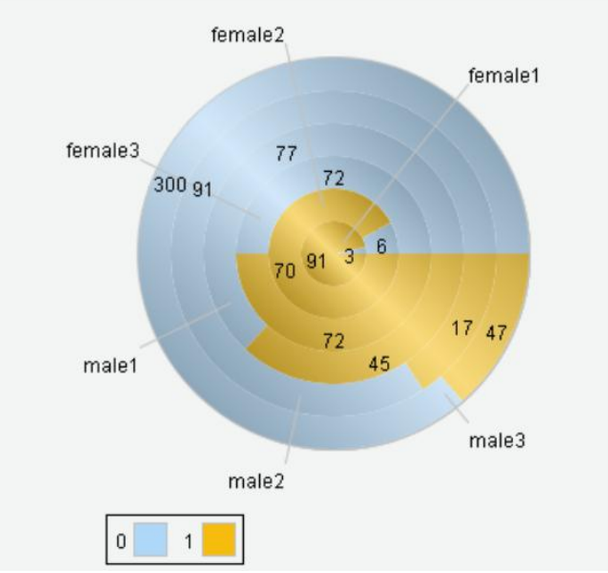
用Sex+Pclass生成交互变量Sex_Pclass

	A	B
1	=file("D:/KDD/titanic.csv").import@qtc(PassengerId,Sex,Pclass)	/导入数据
2	=A1.derive("("+Sex+","+string(Pclass)+")":Sex_Pclass)	/生成Sex_Pclass

A1、A2

Index	PassengerId	Sex	Pclass
1	1	male	3
2	2	female	1
3	3	female	3
4	4	female	1
5	5	male	3
6	6	male	3
7	7	male	1
8	8	male	3
9	9	female	3
10	10	female	2

Index	PassengerId	Sex	Pclass	Sex_Pclass
1	1	male	3	male3
2	2	female	1	female1
3	3	female	3	female3
4	4	female	1	female1
5	5	male	3	male3
6	6	male	3	male3
7	7	male	1	male1
8	8	male	3	male3
9	9	female	3	female3
10	10	female	2	female2



我们可以根据数据的特点，使用变量进行交互，增加衍生变量，然后使用探索的方式方法查看衍生变量是否有用。