1 تحلیل و درک مفاهیم داده

1.1 انتزاع و تعمیم داده ها

1

1. معدل کل (GPA): عددی پیوسته (Interval)
2. حضور در کلاس‌ها: عددی گسسته (Ratio)
3. میزان استفاده از منابع آموزشی مجازی و حضوری: (Ratio)
4. رشته تحصیلی: داده اسمی (Nominal)
5. سطح تحصیلات والدین: داده ترتیبی (Ordinal)
6. وضعیت اقتصادی خانواده: عددی گسسته (Interval/Ratio)
7. میزان تعامل با اساتید و دانشجویان: داده ترتیبی (Ordinal)
8. وجود مشکلات روانی یا استرس: (Nominal)

برای تحلیل احتمال ترک تحصیل دانشجویان نیازمند ویژگی های مختلفی هستیم که بتوان تخمین دقیق تری از ترک تحصیل بدست آوریم پس لازم است که عوامل تاثیر گذار در این ماجرا را بدست آوریم. به عنوان مثال معدل و میزان حضور شخص در کلاس می‌تواند دید خوبی از وضعیت تحصیلی و تمایل به ادامه تحصیل شخص بدهد هر چقد یک دانشجو معدل پایین تر داشته باشد و تمایلی به شرکت در کلاس های درس نداشته باشد احتمال ترک تحصیل آن می‌تواند افزایش یابد.
از طرفی حمایت های تحصیلی از جمله دسترسی به منابع آموزشی و یا وضعیت اقتصادی خانواده در امر ادامه تحصیل بسیار اثر گذار می‌باشد. پس اگر دانشجویی شرایط مالی مطلوبی نداشته باشد و به منابع اموزشی دسترسی نداشته باشد احتمال ترک تحصیل کردن ان حتی در صورتی که معدل بالایی داشته باشد افزایش می‌یابد.همچنین رشته تحصیلی شخص که بیانگر بازار کار آینده دانشجو می‌باشد هم می‌تواند عامل اثر گذاری باشد.
که هر کدام از این داده ها بر ماهیتشان با نوع مختلفی از ویژگی ها تعریف می‌شوند، مثلا معدل دانشجو یک داده پیوسته است و یا حضور در کلا باید به صورت اعداد گسسته تعریف شود. و یا مثلا برای تعامل با دانشجویان و اساتید باید از داده های ترتیبی استفاده کرد(مثلا بد، خوب، خیلی خوب).

2

خیر همه ویژگی ها از اهمیت یکسانی برخوردار نیستند، به عنوان مثال اگر در میان ویژگی ها نام و کدملی دانشجویان قرار داشته باشد این موارد تاثیری در پیش بینی ما نخواهد داشت . و یا اینکه دانشگاه فرد نسبت به وضعیت مالی و نمرات دانشجو اهمیت کمتری داشته باشد.

برای بررسی اینکه کدام ویژگی اهمیت بیشتری دارد هم می‌تواند از دانش خودمان استفاده کنیم مثل مواردی که بالاتر گفته شده است و هم می‌توان از تکنیک هایی همون تحلیل هبستگی استفاده کرد در این روش میزان ارتباط بین ویژگی ها و تغیر هدف (مثل ترک تحصیل) با استفاده از یک ضریب هبستگی یا heatmap بررسی می‌کنیم. و یا می‌توان از روش های انتخاب ویژگی مانند Mutual استفاده کرد.

2.1 اثر نوع ویژگی بر تحلیل داده ها

1

خیر این دو ویژگی را نمی‌توان به یک صورت در تحلیل های آماری در نظر گرفت زیرا نوع داده ها متفاوت است، مثلا درآمد سالانه یک ویژگی عددی پیوسته است که می‌توان بر روی آن عملیات ریاض همچون میانگین و واریانس و یا روش های تحلیل همبستگی اعمال کرد و حتی می‌توان داده های مختلف در درآمد سالانه را باهمدیگر مقایسه کرد.

اما از جهتی کدپستی یک ویژگی گسسته میباشد که صرفا نشان دهنده یک خانه بوده و مقایسه آن با دیگر دادهای کد پستی اطلاعاتی در اختیار ما قرار نمی‌دهد حتی می‌تواند ما را به گمراهی بکشاند از طرفی بر روی این نمونه داده های نمی‌توان عملیات ریاضی همچون واریانس، میانگین و مد اعمال کرد چرا که نتیجه یک عدد می‌شود که هیچ دید آماری به ما نمی‌دهد. صرفا یک عدد بی معنی تولید خواهد شد.

2

برای نرمال سازی داده هایی مثل درآمد سالانه می‌توان از روش هایی همچون Min-Max Scaling و Z-score Normalization استفاده کرد. منتها چون در این روش های محاسبات ریاضی همچون میانگین صورت میگیرد در سوال قبل گفتیم که این نوع روش برای داده هایی مثل کدپستی جوابگو نیست چرا که میانگین گرفتن و یا تفریق و تقسیم چند داده مختلف از نوع کد پستی معنای خاصی به ما نمیدهد و حتی ممکن است در این بین اعداد پیوسته نامفهومی تولید شود.

اما برای اینکه داده ها در بازه بین صفر تا یک واقع شوند می‌توان به هر کدام از کد پستی ها یک عدد در بازه صفرو یک یا به صورت رندم و یا به صورت ترتیبی نسبت داد به عنوان مثال ، کد پستی اول مثلا مقدار صفر را به ان نسبت دهیم کد پستی دوم مقدار 0.001 و کد پستی دوم مقدار 0.002 به همین شکل تا اخر، به عبارتی برای نرمال کردن اینگونه داده های می‌توان از روش های label گذاری استفاده کنیم. که همان طور که مشخص است از روش های محاسباتی و آماری برای نرمال کردن اینگونه داده ها استفاده نشده است.

3.1 نقد معیار توصیفی داده ها

1

وقتی میانگین و میانه برابر باشند اما واریانس بسیار بالا باشد، نشان دهنده یک توزیع متقارن با پراکندگی زیاد است. این به این معنی است که داده ها در اطراف مقدار میانگین توزیع شده اند اما فاصله زیادی از میانگین دارند. این یعنی ممکن است که داده ها از یک توزیع پهن مانند توزیع یکنواخت گسترده یا توزیع نرمال با انحراف معیار بالا باشند.

2

خیر میانگین در این حالت معیار مناسبی برای ما نیست. دلیلیش هم این است که داده ها به دلیل واریانس زیاد از پراکندگی شدیدی برخوردار هستند . پس خود میانگین به تنهایی تصویر دقیق از تمرکز داده ها ارئه نمی‌دهد.

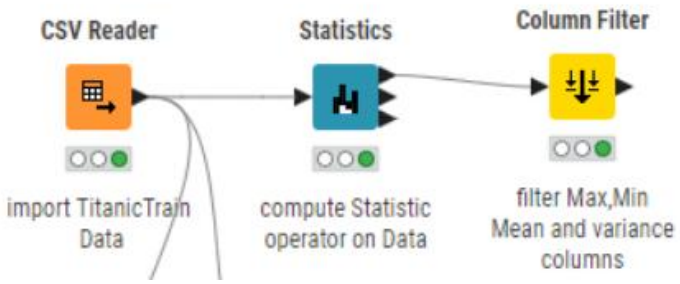
در این حالت می‌توان علاوه بر میانگین از انحراف معیار هم استفاده کرد و یا میتوان از چاک هم استفاده کرد چون داده ها را به باه هایی تقسیم می‌کند و می‌تواند دید آماری بهتری به ما بدهد. از طرفی برای نشان دادن توزیع داده ها میتوان از نمودار هایی همچون هیستوگرام نیز بهره برد.

گزارش بخش عملی

2 تحلیل مجموعه داده تایتانیک

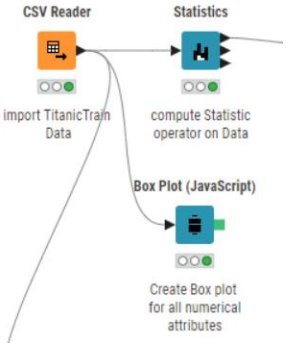
2.1 بصری سازی

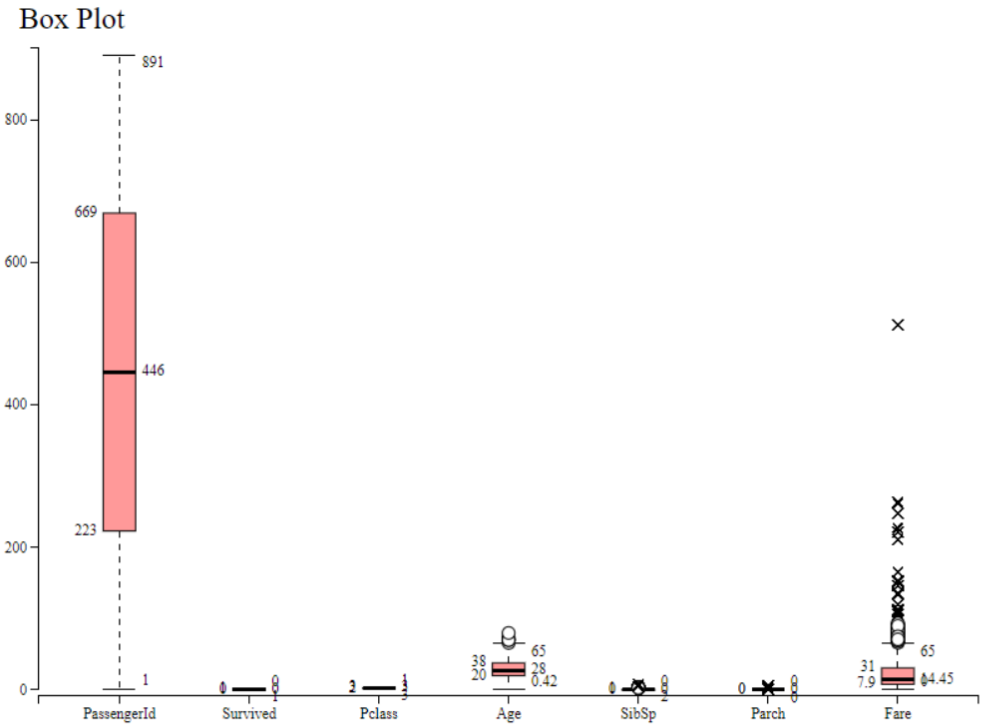
1



#	RowID	Min Number (double)	Max Number (double)	Mean Number (double)	Variance Number (double)
1	Passer	1	891	446	66,231
2	Survive	0	1	0.384	0.237
3	Pclass	1	3	2.309	0.699
4	Age	0.42	80	29.699	211.019
5	SibSp	0	8	0.523	1.216
6	Parch	0	6	0.382	0.65
7	Fare	0	512.329	32.204	2,469.437

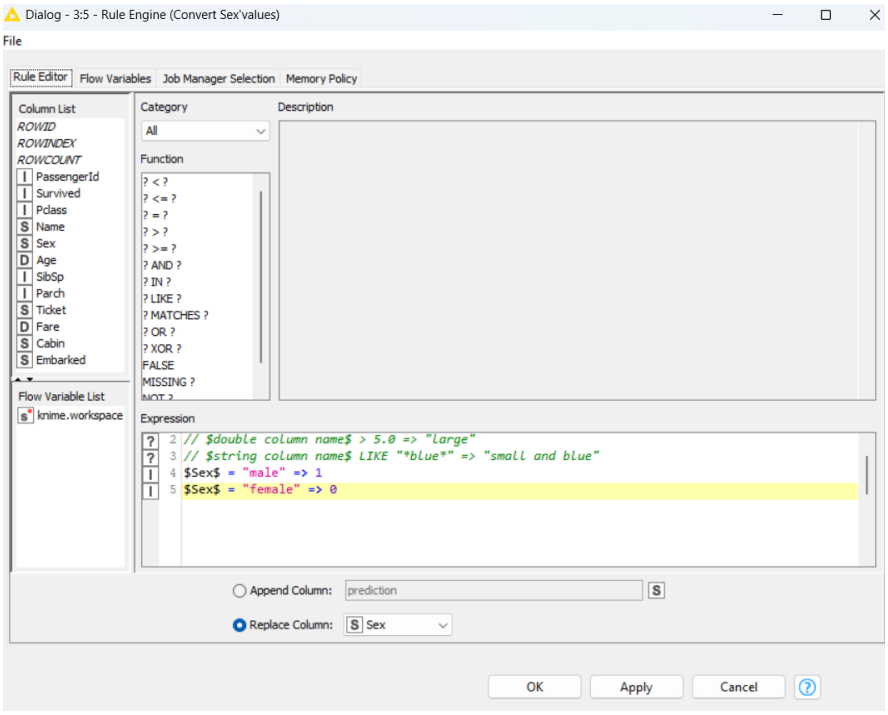
2

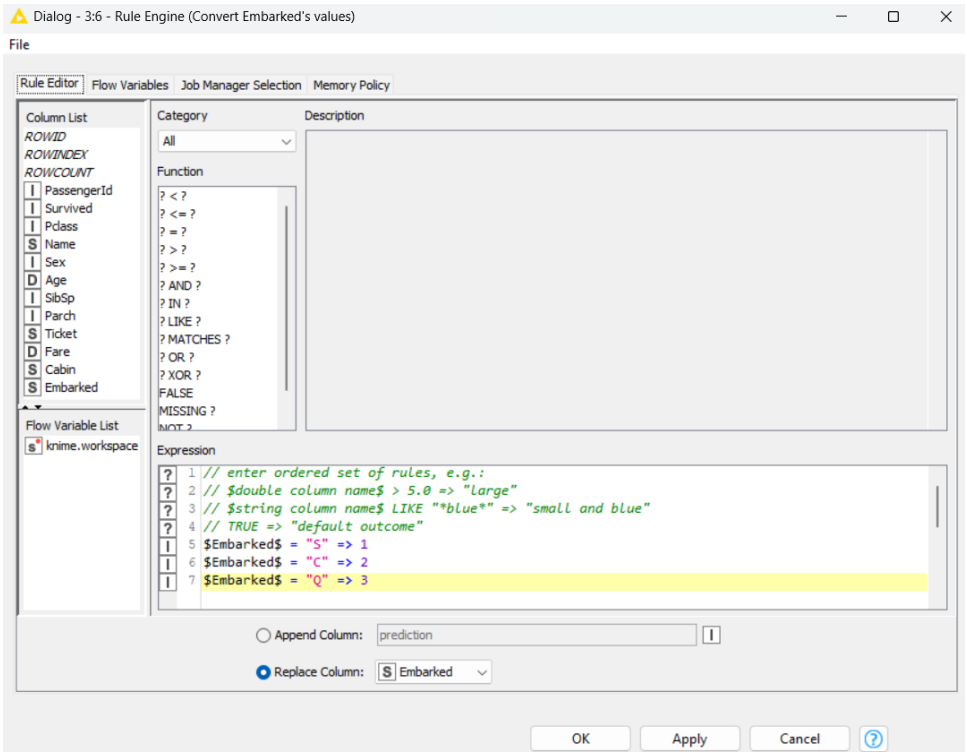




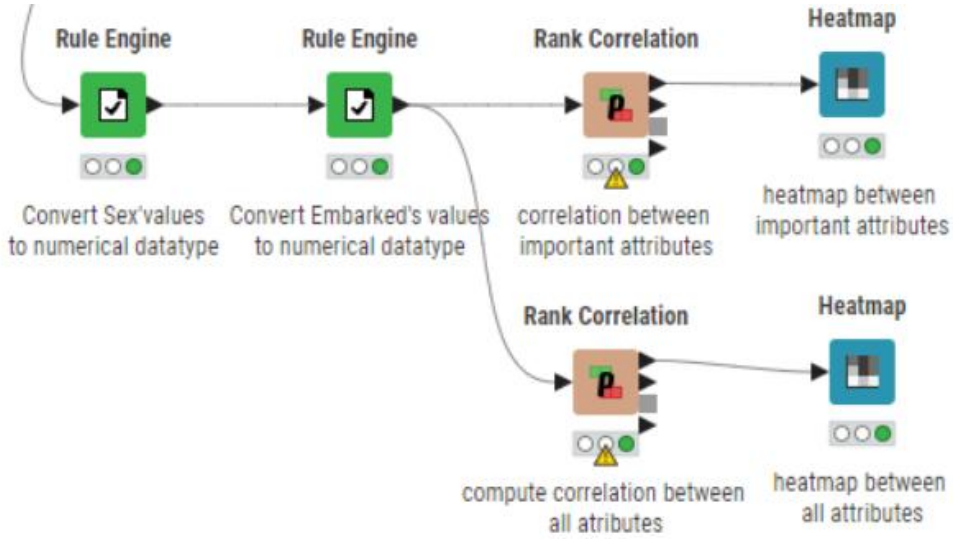
3

برای این سوال ابتدا ویژگی های Sex و Embarked را از حالت string با استفاده از Rule Engine به integer تبدیل میکنیم تا ارزیابی بهتری داشته باشیم .



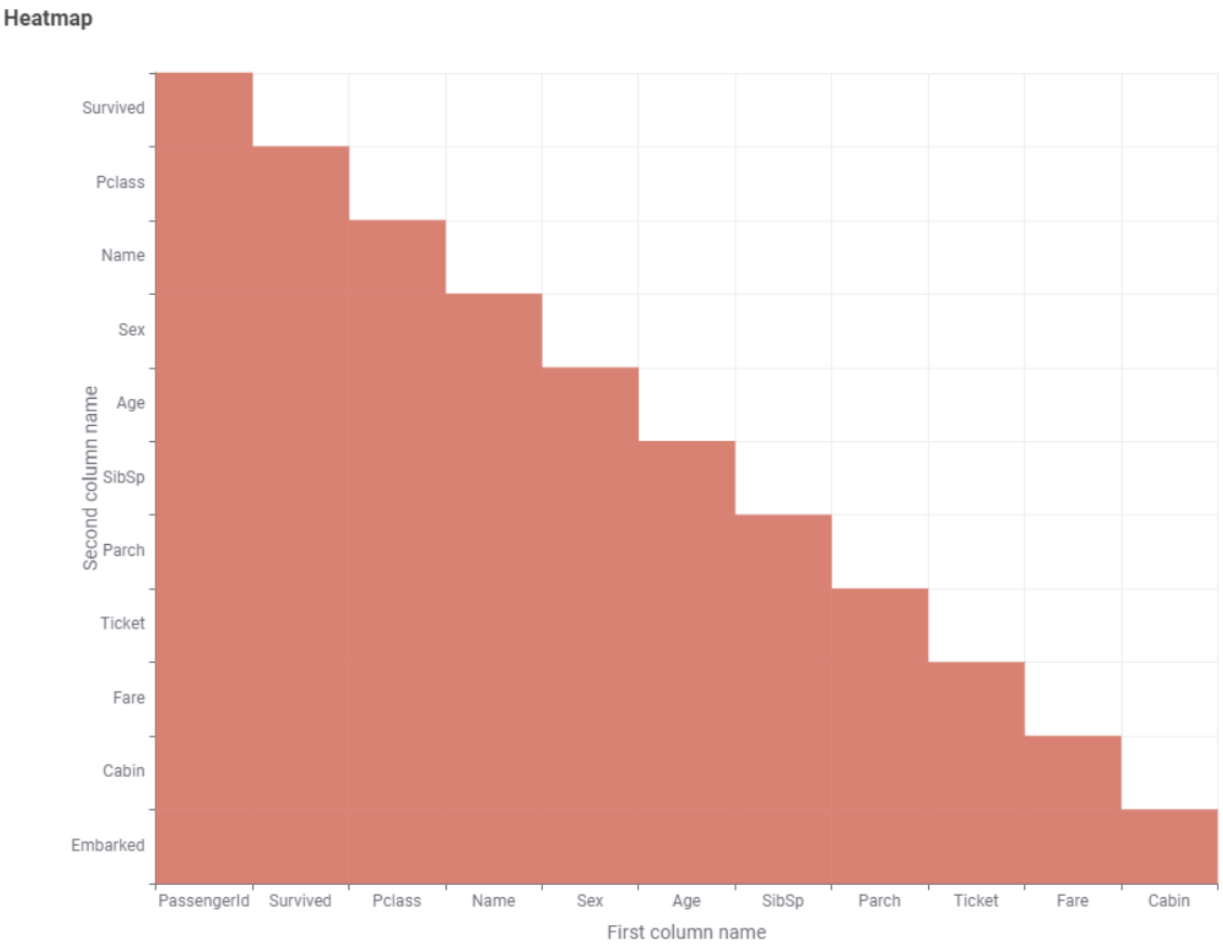


سپس این سوال را در دو حالت بررسی کرده ایم:



حالت اول اینکه ضریب هبستگی همه ویژگی ها حتی اسم و ایدی مسافرین را هم در نظر گرفته ایم و پس از به دست اوردن ضریب هبستگی به نمودارheatmap داده ایم تا متوجه شویم بین کدام دو ویژگی بیشترین همبستگی را داریم.

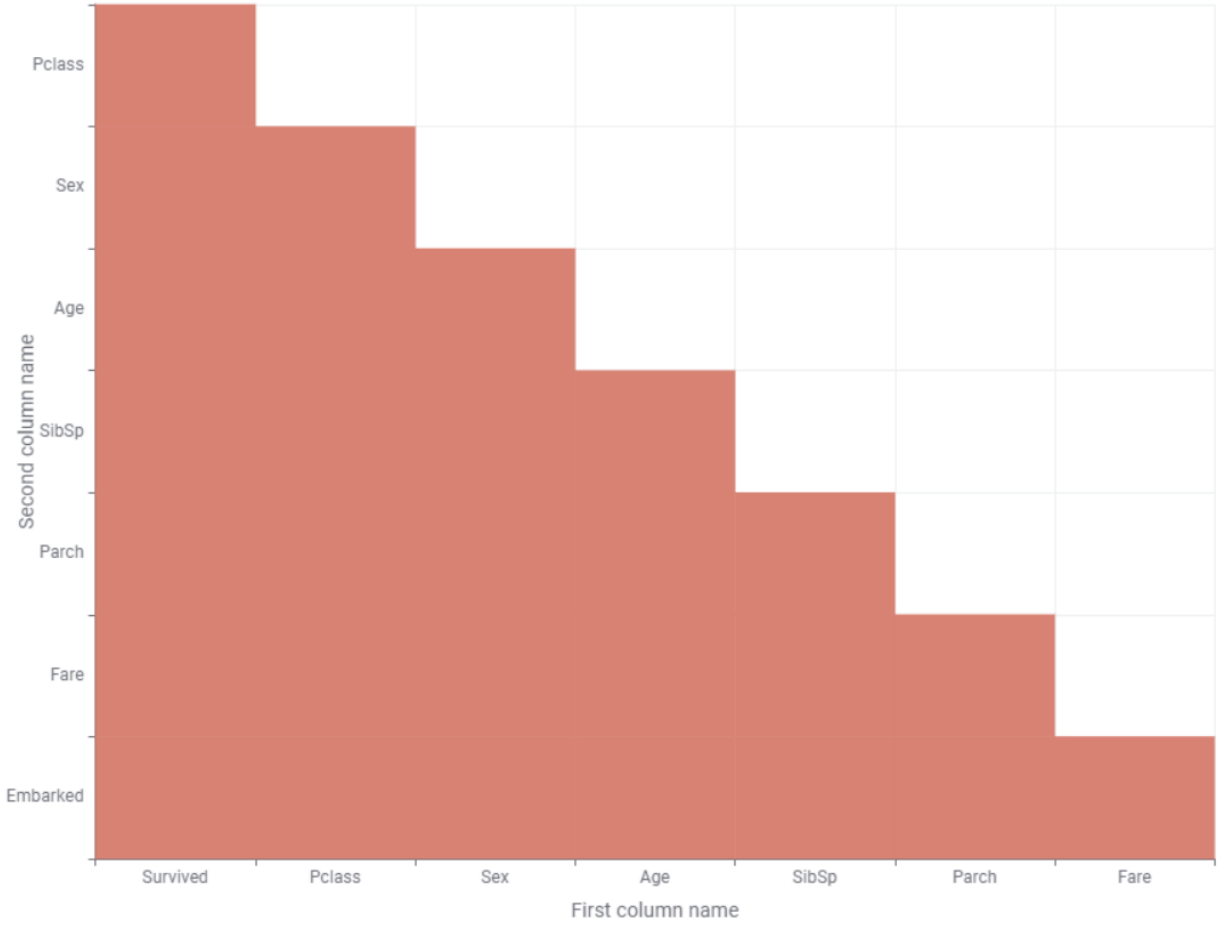
با توجه به نمودار heatmap بدست آمده اگر تمامی attribute ها را لحاظ کنیم ، بیشترین هبستگی میان دو attribute،Survived و PassengerID می‌باشد.



در حالت دوم صرفا ویژگی های مهم را در نظر گرفتیم چرا که داده هایی مثل (Name, PassengerID, Ticket, Cabin) اطلاعات زیادی درباره مسافران در اختیار ما نمیگذارو طبیعتا برای هر مسافر می‌تواند یه مقدار متحصر به فرد داشته باشد،داده کابین هم به این دلیل در نظر نمیگیریم چون مقدار های تعریف نشده زیادی در فیلد های این attribute هست.

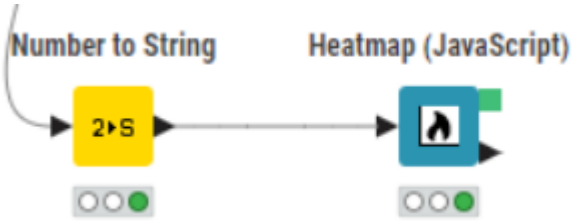
با توجه به نمودار بدست آمده برای حالت دوم بشترین همبستگی میان تو ویژگی Survived و Pclass وجود دارد.

Heatmap

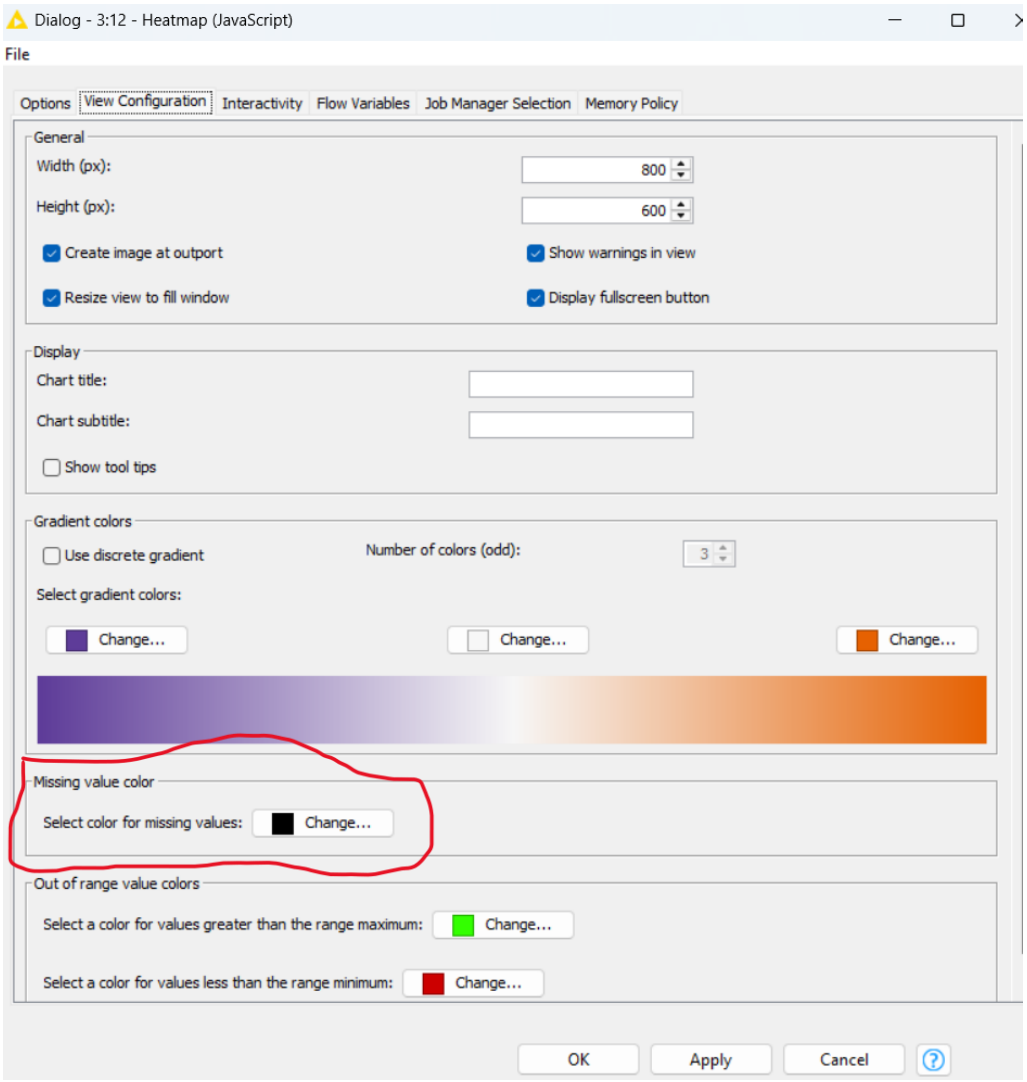


4

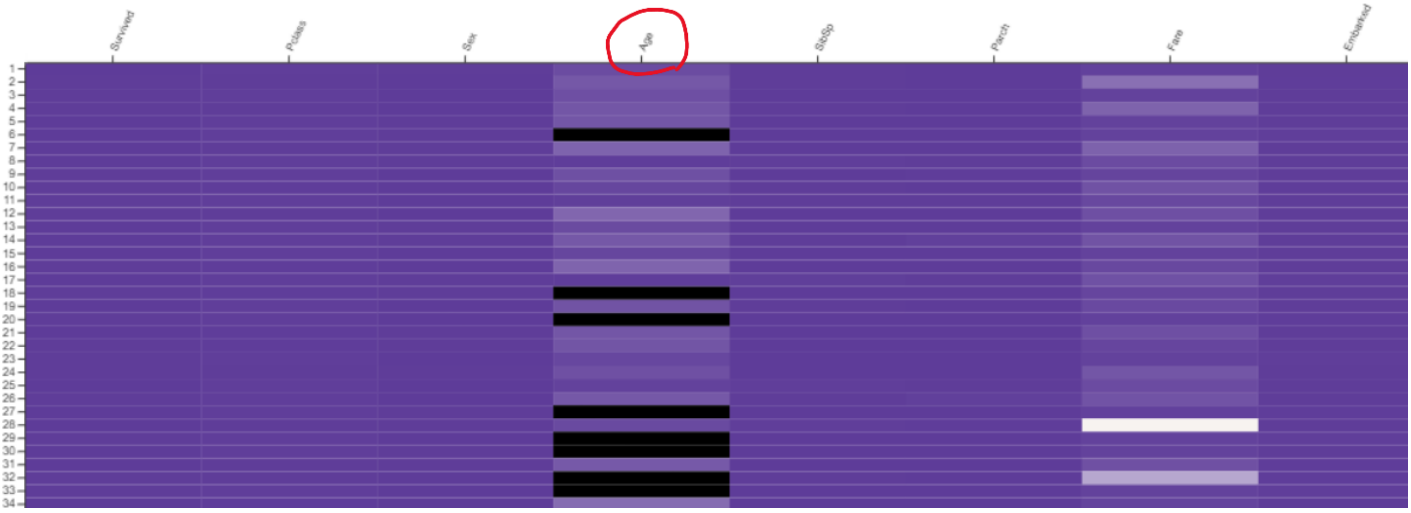
برای اینکه بتوانیم سطر ها را به عنوان PassengerID در نظر بگیریم باید ابتدا این ویژگی را به حالت string تبدیل کنیم. سپس به heatmap بدهیم.

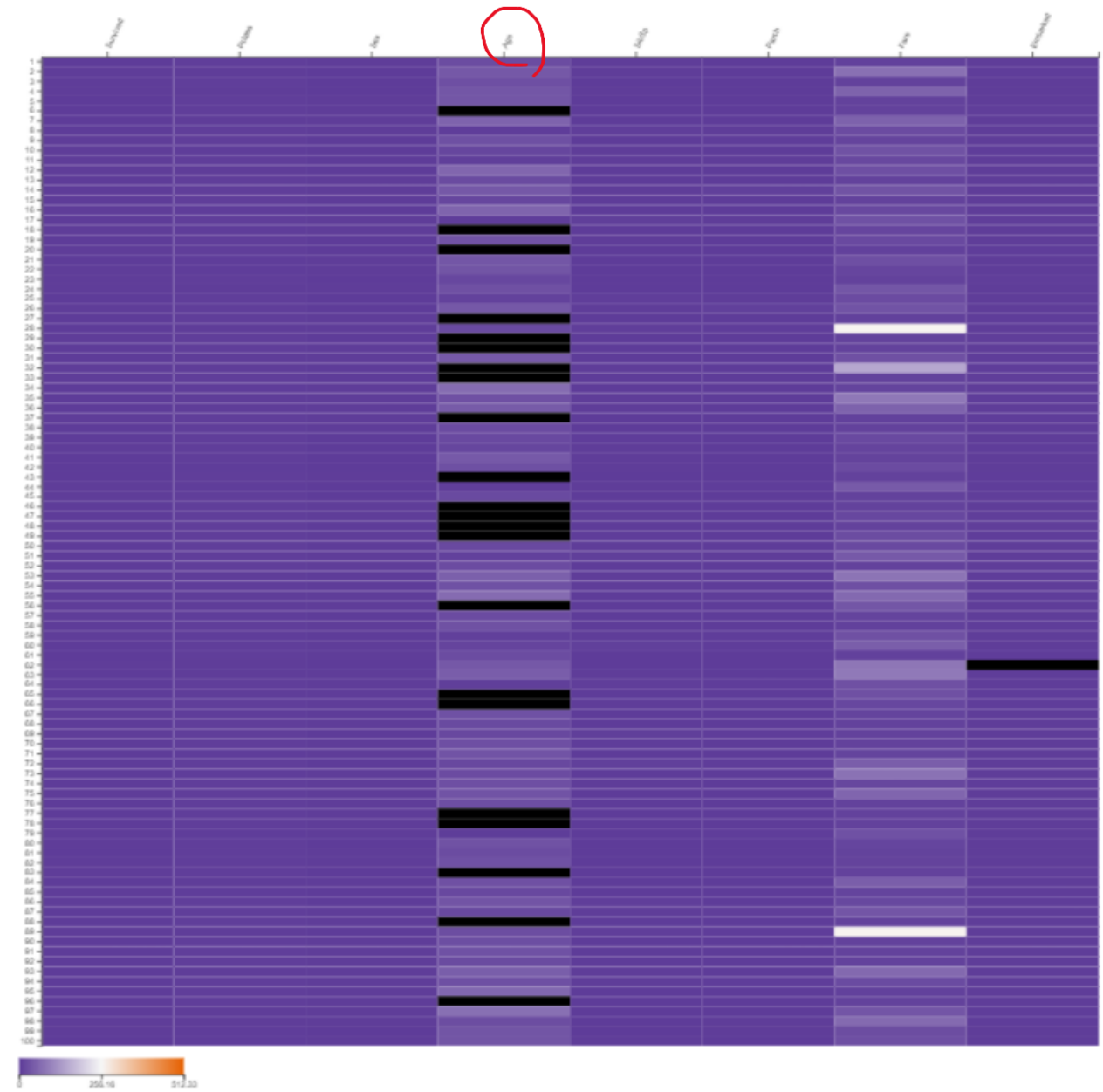


برای نشان دادن مقادیر missing در تنظیمات خود heatmap مشخص شده که این مقادیر باید با رنگ سیاه نشان داده شوند.



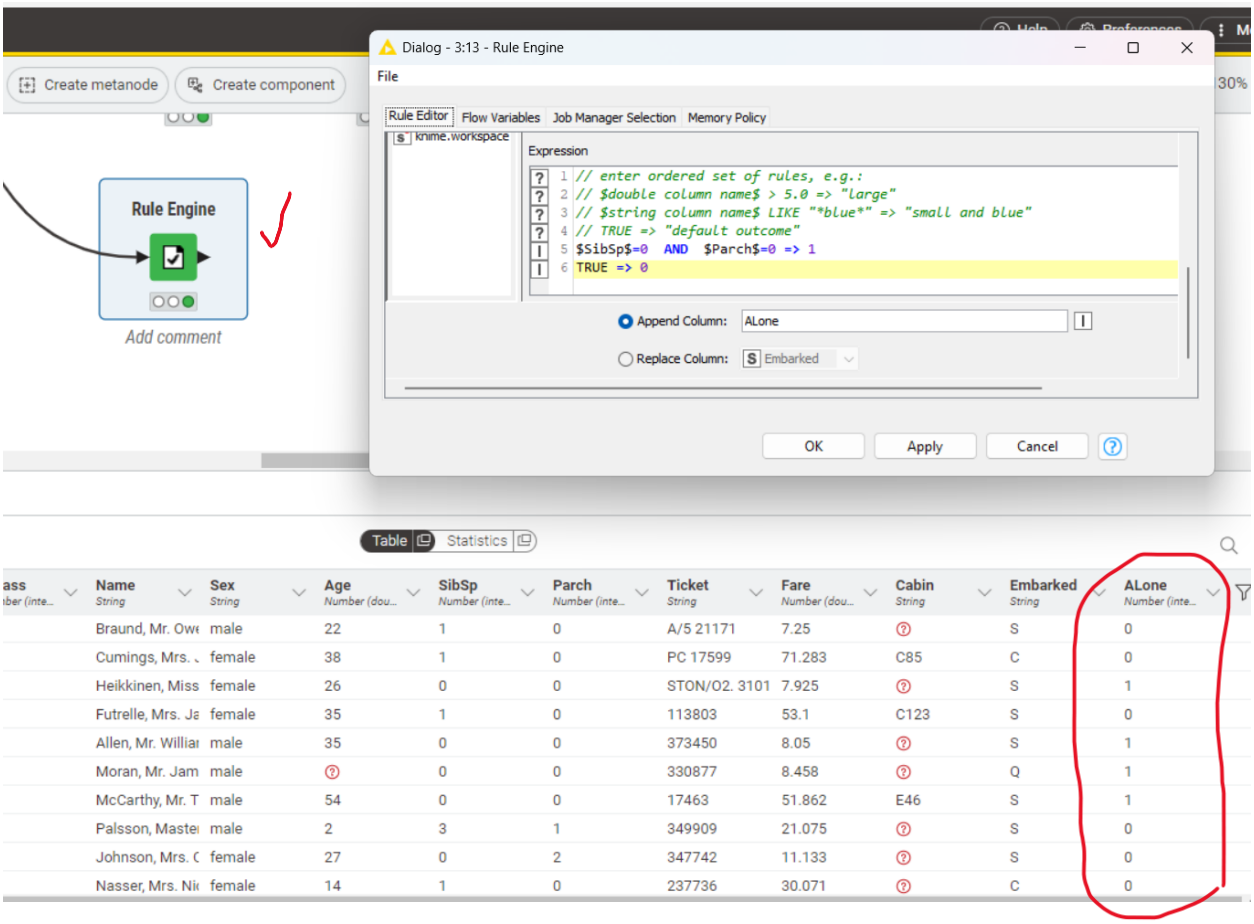
با توجه به خروجی بدست آمده فیلد هایی که مشکی شده اند، missing value هستند و تعداد این مقادیر missing در ستون Age از باقی ستون ها بیشتر می‌باشد.





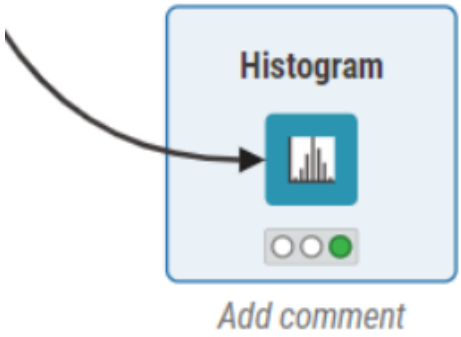
5

برای تعریف یک ویژگی جدید از Rule Engine node استفاده می‌کنیم . و در صورتی که دو ویژگی SibSP و Parch هر دو برابر با صفر بودند یعنی مسافر هیچ وابستگی به خانواده نداشته و تنهایی سفر کرده است.



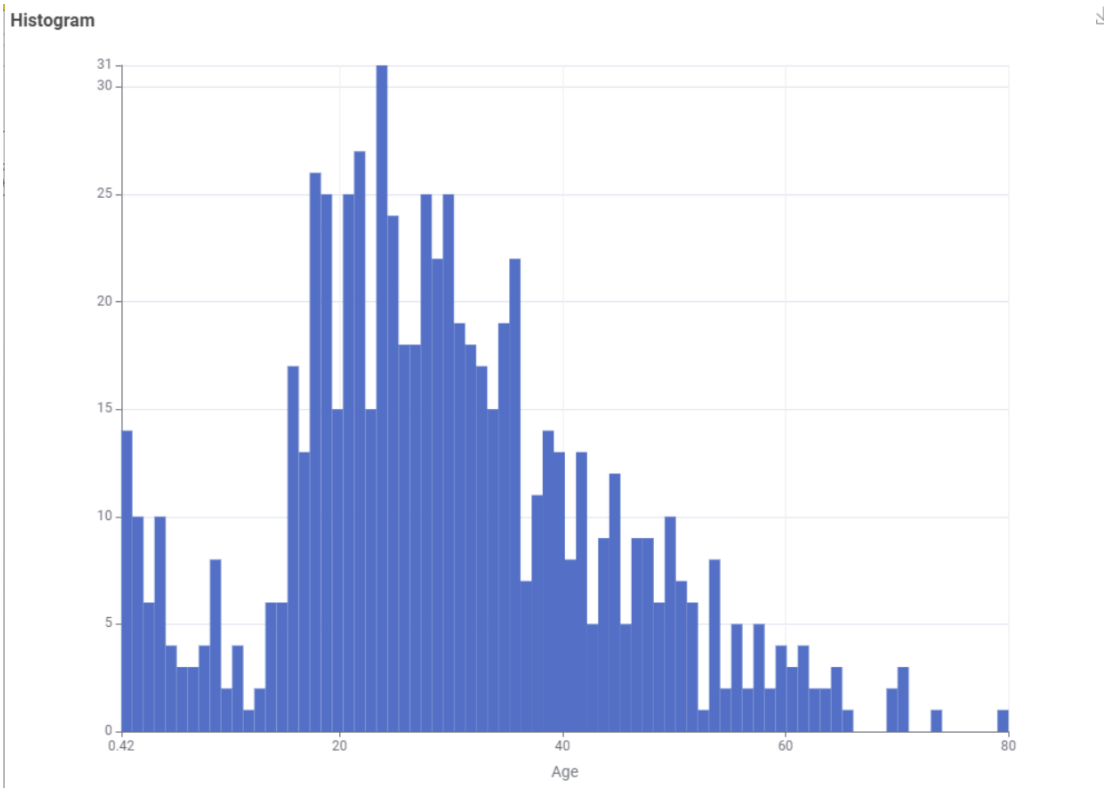
6

برای رسم histogram روی age باید از نودhistogram استفاده کنیم که ورودی های خودش را از داده های csv تایتانیک می‌گیرد.



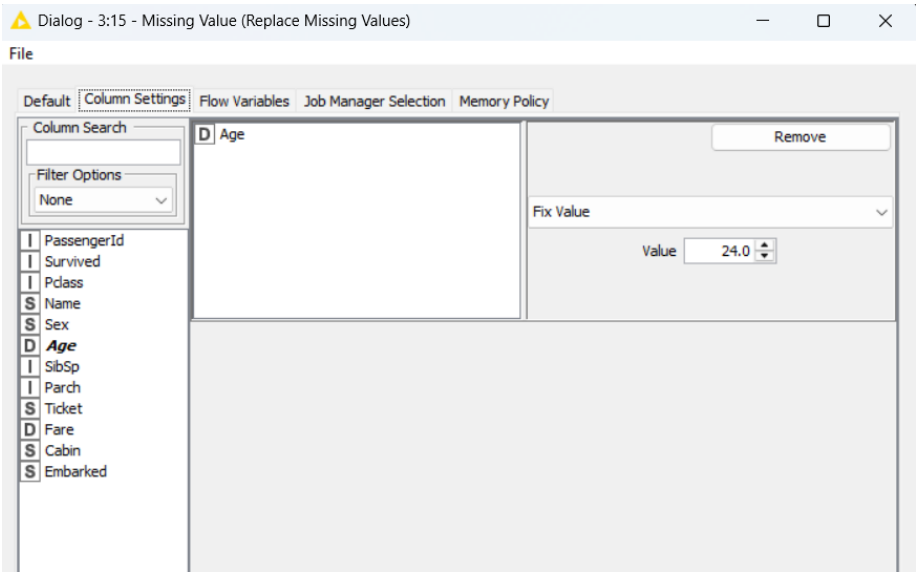
برای رسم هیستوگرام صرفا موارد علامت زده شده در تصویر را تغیر داده ایم و باقی موارد مقدار default خود برنامه هستند

همانطور که در نمودار بدست آمده مشخص است، بیشتر افراد دربازه سنی بین 16 تا 36 سال هستند. تعداد افراد از 50 به بالا نسبت به باقی بازه سنی ها کمتر می‌باشد. بیشترین تعداد افراد با یک سن مشابه 31 نفر هست که سنشان برابر با 24 سال است.



7

برای این سوال لازم است که از نود Missing Value استفاد کنیم. از سوال قبل و با استفاده از نمودار هیستوگرامی که روی Age بدست آوریم ، فهمیدیم که بیشترین سنی که در میان مسافران وجود داشته است برابر با 24 بوده پس مقدار 24 را جایگزین مقادیر Missing در ستون Age می‌کنیم.



مقادیر Age **قبل** از جایگزین کردن Missing valueها

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. '	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr	male		0	0	330877	8.4583		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderco	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, M	female	55	0	0	248706	16		S
17	0	3	Rice, Mast	male	2	4	1	382652	29.125		Q
18	1	2	Williams, M	male		0	0	244373	13		S
19	0	3	Vander Pla	female	31	1	0	345763	18		S
20	1	3	Masselma	female		0	0	2649	7.225		C

مقادیر Age **بعد** از جایگزین کردن Missing valueها

6	Row5	6	0	3	Moran, Mr. Jam	male	24
7	Row6	7	0	1	McCarthy, Mr. T	male	54
8	Row7	8	0	3	Palsson, Maste	male	2
9	Row8	9	1	3	Johnson, Mrs. (	female	27
10	Row9	10	1	2	Nasser, Mrs. Ni	female	14
11	Row10	11	1	3	Sandstrom, Mis	female	4
12	Row11	12	1	1	Bonnell, Miss. E	female	58
13	Row12	13	0	3	Saundercock, M	male	20
14	Row13	14	0	3	Andersson, Mr.	male	39
15	Row14	15	0	3	Vestrom, Miss.	female	14
16	Row15	16	1	2	Hewlett, Mrs. (M	female	55
17	Row16	17	0	3	Rice, Master. Eu	male	2
18	Row17	18	1	2	Williams, Mr. Ch	male	24
19	Row18	19	0	3	Vander Planke,	female	31
20	Row19	20	1	3	Masselmani, Mr	female	24

8

برای پر کردن مقادیر missing در ویژگی کابین روش هایی که می‌توان استفاده کرد به این شکل است:

.1

استفاده از روش های یادگیری ماشین: برای این کار می‌توان از ویژگی هایی مثل Pclass,Fare,Embarked استفاده کرد تا متوجه شویم مسافر در کدام کابین بوده است(مخوصصا چون حرف اول کابین براسا شماره عرشه کشتی می‌باشد می‌توان با این روش ، اینکه مسافر در کدام عرشه می‌باشد را حدس زد). به عنوان مثال اگر Embarked برای مسافری Q باشد و Fare ان از 20 کمتر باشد و Pclass ان 3 باشد احتمالا کابین آن بین حروف F ویا G بوده است.

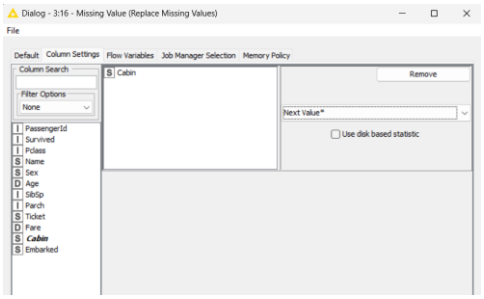
.2

روش دیگر استفاده از نزدیک ترین همسایگی می‌باشد، باید بر اساس اطلاعات سایر مسافرین که اطلاعات آن ها نزدیک به مسافر فعلی مورد بررسی می‌باشد، تخمین بزنیم که شماره کابین مسافر چند بوده است.

.3

یک حالت این است که مقادیر را با مقدار های بعدی و یا قبلی خودشان جایگزین کنیم که این روش میتواند تا حدودی

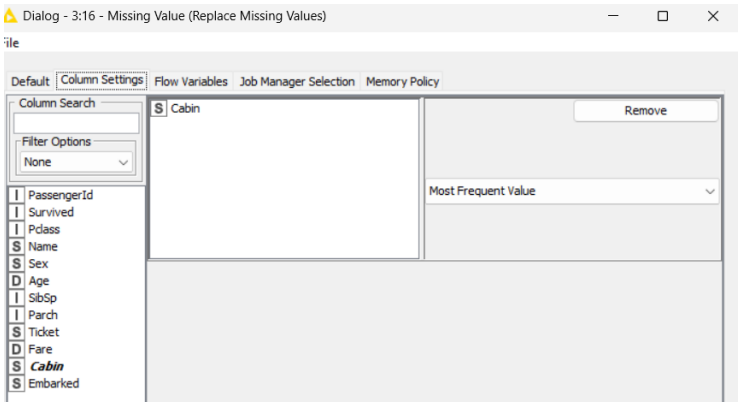
منطقی تر باشد اگر تعداد داده های گم شده خیلی زیاد نباشد که باعث شود اغلب فیلد های این ستون مقداری مشابه بگیرند.



Pclass Number (inte…	Name String	Sex String	Age Number (dou…	SibSp Number (inte…	Parch Number (inte…	Ticket String	Fare Number (dou…	Cabin String
1	Cumings, Mrs. .	female	38	1	0	PC 17599	71.283	C85
3	Heikkinen, Miss	female	26	0	0	STON/O2. 3101	7.925	C123
1	Futrelle, Mrs. Ja	female	35	1	0	113803	53.1	C123
3	Allen, Mr. Williar	male	35	0	0	373450	8.05	E46
3	Moran, Mr. Jam	male	?	0	0	330877	8.458	E46
1	McCarthy, Mr. T	male	54	0	0	17463	51.862	E46
3	Palsson, Maste	male	2	3	1	349909	21.075	G6
3	Johnson, Mrs. C	female	27	0	2	347742	11.133	G6
2	Nasser, Mrs. Ni	female	14	1	0	237736	30.071	G6
3	Sandstrom, Mis	female	4	1	1	PP 9549	16.7	G6
1	Bonnell, Miss. E	female	58	0	0	113783	26.55	C103
3	Saundercock, M	male	20	0	0	A/5. 2151	8.05	D56
3	Andersson, Mr.	male	39	1	5	347082	31.275	D56

.4

حالت بعدی استفاده از روش ها آماری همچون بیشترین مقداری که تکرار شده که البته این روش مناسب نیست به دلیل اینکه نتایج از حالت واقعی خارج می‌شود و نمی‌تواند چند نفر درون یک کابین باشند.



Name String	Sex String	Age Number (dou…	SibSp Number (inte…	Parch Number (inte…	Ticket String	Fare Number (dou…	Cabin String	Embarked String
Braund, Mr. Ow	male	22	1	0	A/5 21171	7.25	G6	S
Cumings, Mrs. .	female	38	1	0	PC 17599	71.283	C85	C
Heikkinen, Miss	female	26	0	0	STON/O2. 3101	7.925	G6	S
Futrelle, Mrs. Ja	female	35	1	0	113803	53.1	C123	S
Allen, Mr. Williar	male	35	0	0	373450	8.05	G6	S
Moran, Mr. Jam	male	?	0	0	330877	8.458	G6	Q
McCarthy, Mr. T	male	54	0	0	17463	51.862	E46	S
Palsson, Maste	male	2	3	1	349909	21.075	G6	S
Johnson, Mrs. C	female	27	0	2	347742	11.133	G6	S
Nasser, Mrs. Ni	female	14	1	0	237736	30.071	G6	C
Sandstrom, Mis	female	4	1	1	PP 9549	16.7	G6	S
Bonnell, Miss. E	female	58	0	0	113783	26.55	C103	S
Saundercock, M	male	20	0	0	A/5. 2151	8.05	G6	S
Andersson, Mr.	male	39	1	5	347082	31.275	G6	S
Vestrom, Miss.	female	14	0	0	350406	7.854	G6	S
Hewlett, Mrs. (M	female	55	0	0	248706	16	G6	S
Rice, Master. Eu	male	2	4	1	382652	29.125	G6	Q

2.1 نقد محدودیت های روش های بصری سازی

1

خیر روش های سنتی مثل هیستوگرام و Scatter plot روش های مناسبی نیستند، چرا که تعداد داده و ویژگی برای این مثلا بسیار زیاد می‌باشد، مثلا از آنجایی که روش Scatter plot می‌تواند نهایتا رابطه بین دو الی سه attributeرا نشان دهد، پس برای مقایسه 50 ویژگی نمی‌توان از این روش استفاده کرد، چرا که مقایسه هر 50 داده باهم ممکن نیست و ثانیا برای مقایسه دودویی هر روش وقت زیاد با دقت پایین حاصل می‌شود. از طرفی به دلیل داشتن 1000نمونه روش هیستوگرام هم روش مناسبی نیست مگر اینکه برای تک تک نمونه های داخل attribute یک bine در نظر نگیریم و صرفا یک بازه را مدنظر قرار دهیم، و اگر بخواهیم هیستوگرام را برای یک attribute بدست آوریم، باید 50 تا نمودار بکشیم که مقدار زیادی می‌باشد.

2

اگر بخواهیم که در فضای دوبعدی نشان دهیم می‌توانیم از تکنیک های کاهش ابعاد مثل (Principal Component Analysis) استفاده کنیم که چند مولفه را بدیل به دو یا سه مولفه می‌کند و در این حالت می‌توانیم نمودار Scatter plot ان را رسم کنیم.

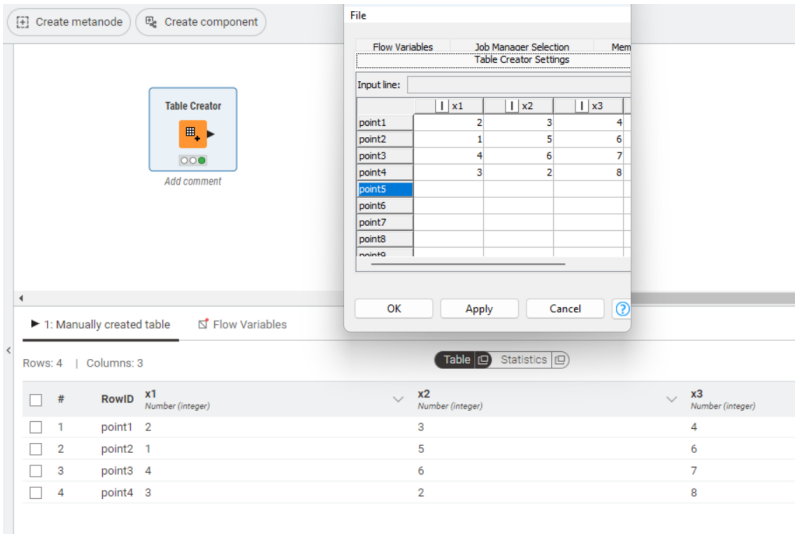
همچنین برای بررسی همبستگی میان ویژگی ها میتوان از نمودار Heatmap هم استفاده کرد. همچنین برای چنین داده هایی می‌توان از نمودار های جدیدتری همچون Parallel Coordinates استفاده کرد، که هر متغیر را میتوان با یک شدت رنگ تعریف کرد.

3 تحلیل فاصله ها در مجموعه داده

3.1

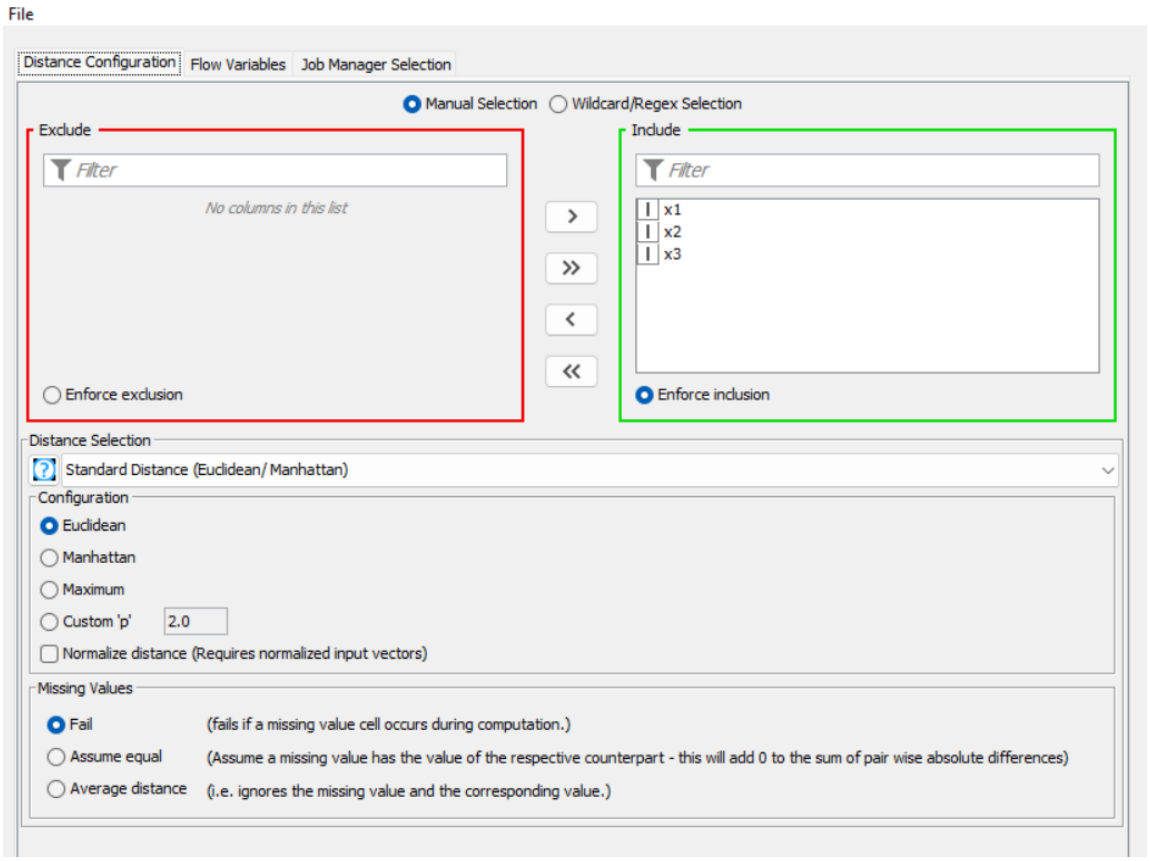
1

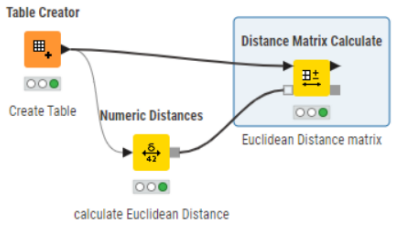
ابتدا باید جدول گفته شده را توسط نود Table Creator بسازیم.

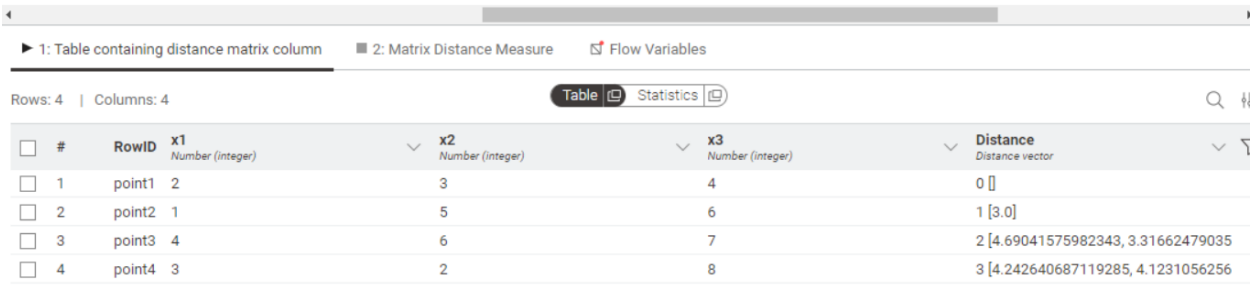


Distance Euclidean .1

برای محاسبه فاصله Euclidean نیاز داریم تا از نود numeric distanceاستفاده کنیم و از نود distance matrix calculate هم برای تشکیل ماتریس فاصله استفاده می‌کنیم.



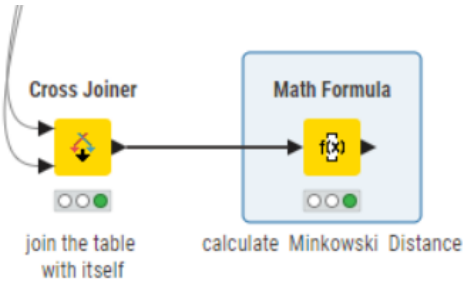


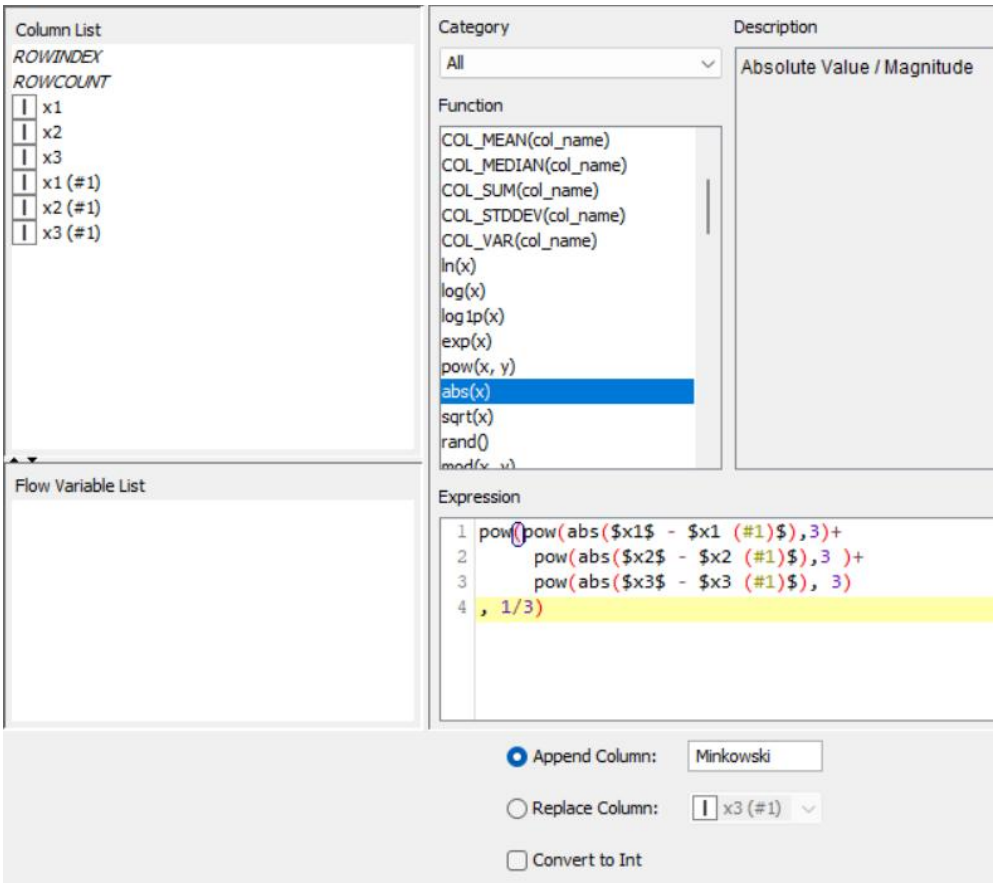


#	RowID	x1 Number (integer)	x2 Number (integer)	x3 Number (integer)	Distance Distance vector
1	point1	2	3	4	0 []
2	point2	1	5	6	1 [3.0]
3	point3	4	6	7	2 [4.6904157598234, 3.31662479035]
4	point4	3	2	8	3 [4.242640687119285, 4.1231056256]

2.Distance Minkowski با r=3

چون به صورت مستقیم نودی وجود ندارد که بتوانیم این فاصله را بدست آوریم، باید به صورت دستی خودمان این روش را پیاده سازی کنیم . پس برای این کار لازم است که ابتدا از نود crosse join استفاده کنیم تا table ساخته شده را با خودش join کنیم تا بتوانیم با توجه به tableبدست آمده ، فاصله هر نقطه با نقاط دیگر را بدست آوریم.پس برای اینکار از فرمول رابطه Minkowski استفاده می‌کنیم.



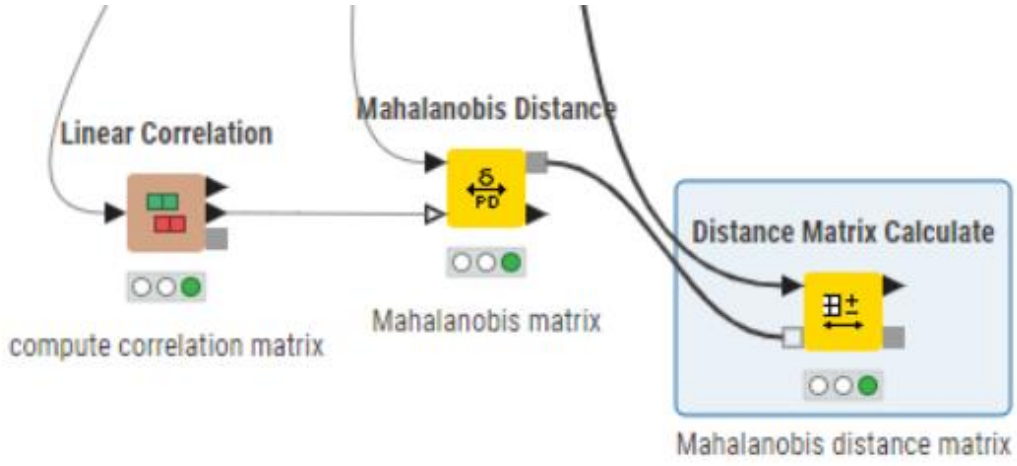


با توجه به شکل زیر هر کدام از سطر های ستون Minkowski به ترتیب می‌شود: فاصله بین نقطه اول با خودش، فاصله نقطه اول با نقطه دوم ، فاصله نقطه اول با نقطه سوم و فاصله نقطه اول با نقطه چهارم و بعد از آن فاصله نقطه دوم تا خودش و... .

#	RowID	x1 Number (integer)	x2 Number (integer)	x3 Number (integer)	x1 (#1) Number (integer)	x2 (#1) Number (integer)	x3 (#1) Number (integer)	Minkowski Number (double)
1	point1_	2	3	4	2	3	4	0
2	point1_	2	3	4	1	5	6	2.571
3	point1_	2	3	4	4	6	7	3.958
4	point1_	2	3	4	3	2	8	4.041
5	point2_	1	5	6	2	3	4	2.571
6	point2_	1	5	6	1	5	6	0
7	point2_	1	5	6	4	6	7	3.072
8	point2_	1	5	6	3	2	8	3.503
9	point3_	4	6	7	2	3	4	3.958
10	point3_	4	6	7	1	5	6	3.072
11	point3_	4	6	7	4	6	7	0
12	point3_	4	6	7	3	2	8	4.041
13	point4_	3	2	8	2	3	4	4.041
14	point4_	3	2	8	1	5	6	3.503
15	point4_	3	2	8	4	6	7	4.041
16	point4_	3	2	8	3	2	8	0

Distance Mahalanobis.2

در ابتدا باید correlation را حساب کنیم و بعد با استفاده از node های Mahalanobis و Ditance matrix calculate فاصله بین نقاط را بدست آوریم.



نهایتا خروجی به شکل زیر می‌باشد:

#	RowID	x1 Number (integer)	x2 Number (integer)	x3 Number (integer)	Distance Distance vector
1	point1	2	3	4	0 []
2	point2	1	5	6	1 [3.978684573661562]
3	point3	4	6	7	2 [4.2426637922478925, 3.122781860
4	point4	3	2	8	3 [4.284686258373692, 3.9579678054

3.1

Euclidean Distance

ازنقطه	به نقطه1	به نقطه2	به نقطه3	به نقطه 4
نقطه 1	0	3	4.69041575982343	4.242640687119285
نقطه2	3	0	3.3166247903554	
نقطه3	4.69041575982343	3.3166247903554	0	
نقطه4	4.2426406871192	4.12310562561766	4.242640687119285	0

Minkowski Distance

ازنقطه	به نقطه1	به نقطه2	به نقطه3	به نقطه 4
نقطه 1	0	2.571	3.958	4.041
نقطه2	2.571	0	3.072	3.503
نقطه3	3.958	3.072	0	4.041
نقطه4	4.041	3.503	4.041	0

Mahalanobis Distance

ازنقطه	به نقطه1	به نقطه2	به نقطه3	به نقطه 4
نقطه 1	0	3.978684573661562	4.2426637922478925	4.284686258373692
نقطه2	3.978684573661562	0	3.12278186072464	3.9579678054560534
نقطه3	4.2426637922478925	3.12278186072464	0	4.280898519276987
نقطه4	4.284686258373692	3.9579678054560534	4.280898519276987	0

اگر همبستگی داده ها بهم کم باشد میتوان از دو ویژگی Minkowski و Euclidean استفاده کرد، اما در اصل Minkowski میتواند فاصله های بیشتر بین داده ها را به خوبی نشان بدهد، پس میتواند معیار

بهتری برای نشان دادن تفاوت های میان داده ها باشد ،اما اگر داده ها با توجه به شرایط گفته شده در بالا همبستگی داشته باشند بایکدیگر این روش بهتر استMahalanobis .

3.3 با توجه به چهار زمینه داده اي بالا:

1

در این حالت بهتر است که از فاصله Minkowski با r=1 استفاده کنیم، زیرا که اگر داده ها همبسته نباشند ممکن است که Mahalanobis نتایج غلط باشد (اگر همبسته باشند گزینه خوبی هست) و فاصله اقلیدسی هم برای داده های گسسته جواب بی معنی ممکن است بدهد.

2

فاصله Mahalanobis بهترین گزینه است چون رابطه همبستگی ویژگی ها را در نظر می‌گیرد و داده ها را در فضایی نرمال شده مقایسه می‌کند.

3

فاصله اقلیدسی و Minkowski نیاز به نرمال سازی داده ها دارد. اما Mahalanobis به دلیل در نظر گرفتن کواریانس، نیاز به نرمال سازی ندارد و داده ها را متعادل می‌کند.

4

اگر داده ها به شدت غیر نرمال باشند فاصله Minkowski خیلی بهتر از دو فاصله دیگر نتایج را برمی‌گرداند.