آموزش

شناسایی داده های پَرت با SPSS

تدوین: رامین کریمی
www.kharazmi-statistics.ir

داده های پرت چیست؟

همیشه باید داده هایی (اطلاعاتی) که وارد برنامه هایی مثل اِکسل یا SPSS می کنیم را وارسی و بازبینی کنیم. همواره احتمال دارد که در داده ها با مقادیر غیر عادی مواجه شویم. موارد غیرعادی می تواند شامل مقادیر گمشده و مقادیر پَرت (دور افتاده) باشد. همواره قبل از انجام هرگونه تحلیل آماری بر روی داده ها، می بایست چاره ای در مورد مقادیر پرت بیندیشیم.

افرادی که اندازه‌های انتهایی یا غیرمعمول در یک متغیر واحد (تک متغیری) یا در ترکیبی از متغیرها (چندمتغیری) دارند، دور افتاده یا پرت (Missing Values) نامیده می شوند. داده های پرت اغلب سه یا بیش از سه واحد انحراف معیار (SD ۳±) از میانگین مربوط به خودشان فاصله دارند که از مشکلات احتمالی در ابزار اندازه گیری، شیوه ثبت یا ضبط پاسخ ها یا عضویت شرکت کنندگان در جامعه ای که فرض می شود از آن نمونه گیری شده است، ناشی می شود. حضور داده های پرت می تواند نتایج تحلیل را به گونه ای نامطلوب تحت تاثیر قرار دهد(تحریف کند). به همین دلیل بیشتر متخصصان آمار پیشنهاد می کنند که اندازه های پرت قبل از تحلیل داده ها باید حذف شوند(میزر، گامست و گارینو،۱۳۹۱:۲۳۶).

انواع داده های پرت

داده های پرت را می توان در دو دسته داده های پرت تک متغیری و داده های پرت چندمتغیری تقسیم کرد:.

الف) داده های پرت تک متغیری

داده های پرت تک متغیری در یک متغیر وجود دارند. به عنوان مثال وقتی که در یک تحقیق دانشجویی در زمینه میزان رضایت مردم از عملکرد شهرداری؛ ما در متغیر سن افراد با عدد ۱۸۰ روبرو می شویم!، به احتمال زیاد با داده پرت مواجه شده ایم. چرا که می دانیم احتمال وجود فردی با چنین سن و سالی بسیار بعید است! و یا وقتی که در متغیر درآمد، شخصی درآمد ماهانه خود را از یک کار تمام وقت ۲۵ هزار تومان اعلام می کند و یا وقتی که در پاسخ سوالی که از فرد می پرسیم تا چه اندازه به آینده امیدوار است و او بایست میزان رضایت خود را از عدد ۱ (به معنای خیلی کم) تا عدد ۵

(به معنی خیلی زیاد) اعلام کند، در فایل داده ها با عدد ۶ روبرو می شویم! (به دلیل اشتباه در ورود داده)، همگی نشان از وجود داده های پرت تک متغیری دارد که نخست می بایست آن ها را شناسایی کرد و سپس در مورد آن ها چاره ای اندیشید. البته زمانی که با متغیرهای کیفی (اسمی و ترتیبی) سروکار داریم گاهی با مقادیری در داده ها روبرو می شویم که داده پرت محسوب نمی شوند اما مقادیری هستند که به اشتباه وارد شده اند و باید حذف شوند. مثلا در متغیر جنس، اگر ما زنان را با کد ۱ و مردان را با کد ۲ معرفی کرده باشیم و در این حال با عدد ۱.۵ در داده ها مواجه شویم با داده پرت مواجه نیستیم اما با داده های اشتباه مواجه شده ایم و بایست آن ها را شناسایی کرده و آن ها را حذف یا اصلاح نماییم.

شناسایی داده های پرت تک متغیری

برای شناسایی داده های پرت تک متغیری می بایست از جدول فراوانی و نمودار جعبه ای استفاده کرد. از جدول فراوانی برای شناسایی داده های پرت در متغیرهای اسمی و ترتیبی استفاده می کنیم و از نمودار جعبه ای برای شناسایی داده های پرت در متغیرهای فاصله ای/ نسبی. البته از جدول فراوانی هم می توان برای شناسایی داده های پرت در متغیرهای فاصله ای/ نسبی استفاده کرد ولی نمودار جعبه ای برتری دارد و آسان تر است.

۱) جداول فراوانی

از جداول فراوانی برای کشف مقادیر پرت تک متغیری در متغیرهای اسمی و ترتیبی استفاده می کنیم. متغیرهایی مثل جنس، وضعیت تاهل، قومیت، تحصیلات و درآمد (هر دو به صورت چندگزینه ای و ترتیبی سنجیده شده باشند، مثلا تحصیلات در قالب سوالات دیپلم، فوق دیپلم، لیسانس و... سنجیده شده باشد) و یا تمام سوالاتی که در قالب طیف لیکرت سنجیده شده باشند. یعنی سوالاتی که پاسخ های آنان معمولا ۳ تا ۷ گزینه دارد و پاسخ هایی مثل کاملا موافقم تا کاملا مخالفم، اصلا تا همیشه و خیلی کم تا خیلی زیاد را در بر می گیرد.

مثال

در یک تحقیق (فرضی) از دانشجویان دختر و پسر دانشگاه شهید بهشتی خواسته شد تا میزان رضایت خودشان از عملکرد ریاست دانشگاه را اعلام کنند. بر این اساس از دانشجویان تعدادی سوال پرسیده شد که دو سوال آن عبارت بود از جنس دانشجویان و میزان رضایت‌شان از عملکرد ریاست دانشگاه. جنس دانشجویان شامل دو جنس (دختر کد ۱، و پسر کد ۲) و میزان رضایت در طیف لیکرت ۵ گزینه ای (خیلی کم کد ۱، کم کد ۲ ، متوسط کد ۳، زیاد کد ۴ و خیلی زیاد کد ۵) سنجیده شد. همانطور که مشاهده می شود ما هنگام ورود اطلاعات مربوط به جنس افراد به دانشجویان دختر کد یا عدد ۱ و به دانشجویان پسر کد ۲ داده ایم و در فایل داده ها و خروجی (برونداد) مربوط به آن، تنها باید عدد ۱ و عدد ۲ مشاهده کنیم. در مورد متغیر میزان رضایت هم تنها باید اعداد ۱، ۲، ۳، ۴ و ۵ را مشاهده کنیم و نباید اعداد دیگری را (مثلا ۶، ۱.۵، ۲۰) مشاهده کنیم.

نحوه اجرای دستور فراوانی در SPSS

دستور فراوانی را اجرا می کنیم:

Analyze ---> Descriptive Statistics ---> Frequencies

درکادر مربوطه متغیرهای جنس و رضایت را وارد کادر Variables (متغیرها) می کنیم و گزینه Ok را انتخاب می کنیم.

نتایج جدول فراوانی دو متغیر جنس و میزان رضایت در ادامه ارائه شده است.

در جدول فراوانی جنس افراد مقادیر پرت مشاهده نمی شود، چرا که تنها دو کد یا طبقه ۱ و ۲ (دختر و پسر) وجود دارند. توجه شود که داده های گمشده (Missing) جزء داده های پرت به حساب نمی آیند. ما در فایل داده ها مقادیر گمشده را با عدد ۹ نشان داده ایم و در فایل خروجی اعداد گمشده با عدد ۹ ظاهر شده اند. به غیر از اعداد ۱ و ۲ و مقادیر گمشده، عدد دیگری در فایل خروجی جنس دانشجویان دیده نمی شود و بدین معناست که در متغیر جنس دانشجویان داده پرت وجود ندارد.

اما در متغیر میزان رضایت ما با اعدادی غیر از ۱ ، ۲، ۳، ۴ و ۵ مواجه ایم و این اعداد مقادیر گمشده هم نیستند و نشان می دهد که دو مقدار پرت در داده ها وجود دارد (۱.۳ و ۲۲) که می بایست در فایل داده ها شناسایی و حذف شود. چون پاسخگویان تنها می توانستند یکی از اعداد ۱، ۲، ۳، ۴ و ۵ را انتخاب کنند در نتیجه اعداد دیگری که وجود دارند (۱.۳ و ۲۲) مقادیر پرت حساب می شوند و باید از تحلیل حذف شوند.

جنس

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	133	66.5	66.8	66.8
	2	66	33.0	33.2	100.0
	Total	199	99.5	100.0	
Missing	9	1	.5		
Total		200	100.0		

میزان رضایت

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.0	94	47.0	47.0	47.0
	1.3	1	.5	.5	47.5
	2.0	48	24.0	24.0	71.5
	3.0	37	18.5	18.5	90.0
	4.0	12	6.0	6.0	96.0
	5.0	7	3.5	3.5	99.5
	22.0	1	.5	.5	100.0
	Total	200	100.0	100.0	

۲) نمودار جعبه ای

اگر متغیرهایی که سنجیدیم از نوع متغیرهای فاصله ای/ نسبی باشند هم می توان از جداول فراوانی استفاده کرد و هم از نمودار جعبه ای. البته پیشنهاد می شود از نمودار جعبه ای استفاده شود. زیرا در خود نمودار جعبه ای داده های پرت شناسایی شده و شماره موردی (پاسخگویی) که دارای داده پرت در پرسشنامه است در نمودار مشخص است. همچنین مبنای انتخاب داده پرت در نمودار جعبه ای، داشتن فاصله ای به اندازه حداقل ۳± واحد انحراف استاندارد و بیشتر با میانگین است که به صورت خودکار توسط خود برنامه انجام می شود و در نمودار جعبه ای نشان داده می شود.

مثال

در پرسشنامه رضایت دانشجویان علاوه بر سوالات جنس و میزان رضایت، سن و معدل دانشجویان هم پرسیده شد و دانشجویان سن و معدل خودشان را در پرسشنامه نوشتند که در نتیجه سن و معدل دانشجویان به صورت طیف فاصله ای/نسبی سنجیده شد. دو نمودار جعبه ای سن و معدل دانشجویان در ادامه آورده شده است.

همانطور که در نمودار سن افراد مشاهده می شود دو مورد دارای مقادیر پرت هستند. شماره این دو مورد در نمودار با علامت ستاره یا دایره کوچک مشخص شده است. افراد شماره ۱۲۲ و ۱۲۹ دارای مقادیر پرت هستند و به احتمال زیاد باید از داده ها حذف شوند. سن این افراد به ترتیب ۷۸ و ۷۵ سال است. با توجه به این که احتمال وجود دانشجویانی با چنین سنی بسیار بعید است سن این افراد باید از تحلیل حذف شود.

در متغیر معدل سه داده پرت وجود دارد که شامل موردهای ۹۵، ۱۹۵ و ۱ می شود. معدل این افراد به ترتیب ۱۲، ۱۲ و ۷.۲۱ است. تصمیم در مورد حذف این داده ها با پژوهشگر است. معدل این افراد حداقل ۳± واحد انحراف استاندارد با معدل کل دانشجویان فاصله دارد. معدل کل دانشجویان برابر با ۱۶.۵۰ و انحراف استاندارد ۱.۷۵ است. به نظر می رسد با این که معدل ۱۲، معدل پایینی است و احتمال این که دانشجویی معدل ۱۲ داشته باشد اندک است اما همیشه چنین دانشجویانی وجود داشته اند و فرض وجود چنین معدل هایی محتمل است. بنابراین موردهای ۹۵ و ۱۹۵ باقی مانده ولی مورد ۱ که دارای معدل ۷.۲۱ است حذف می شود، چون دارای معدل خیلی پایینی است.

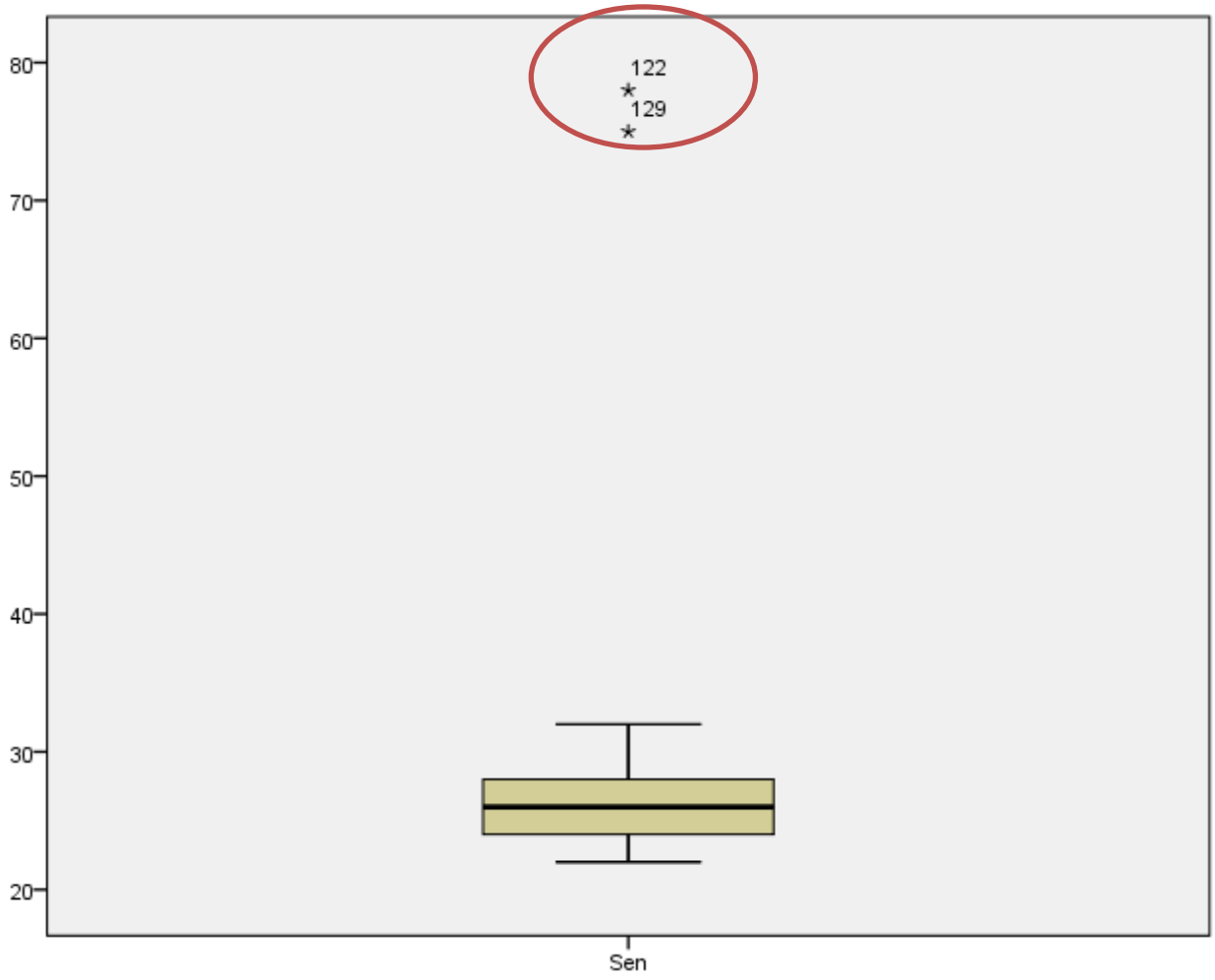
نحوه اجرای دستور نمودار جعبه ای در SPSS

دستور اجرای نمودار جعبه ای در دستور Explore است. مراحل زیر را دنبال می کنیم

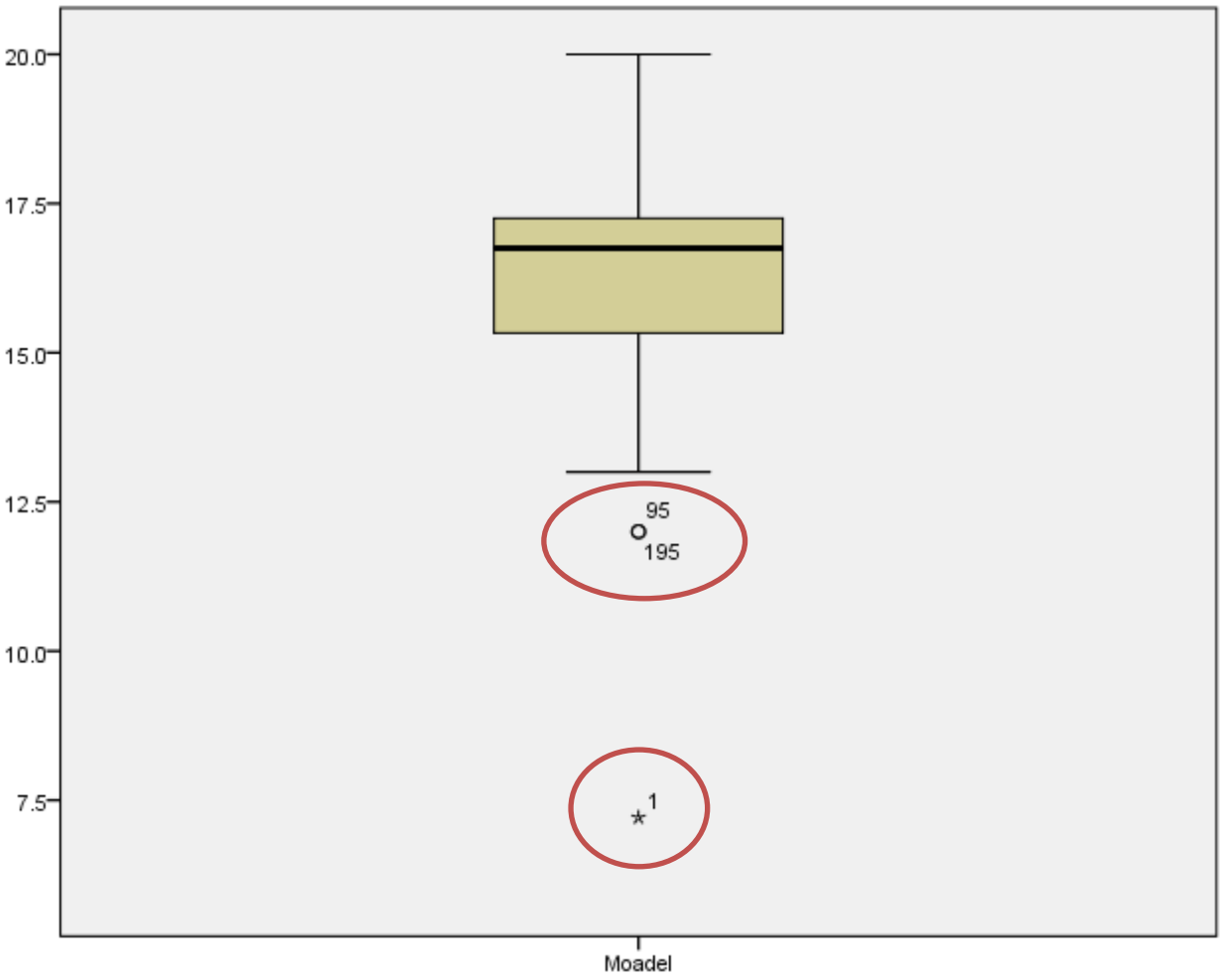
Analyze ---> Descriptive Statistics ---> Explore

متغیر مدنظر را وارد کادر Dependent List می کنیم.

در کادر Explore گزینه Plots را انتخاب کرده و گزینه Normality plots with tests را فعال می کنیم.



نمودار جعبه ای سن



نمودار جعبه ای معدل

ب) داده های پرت چند متغیری

یک روش عینی‌تر برای ارزیابی وجود داده‌های پرت چندمتغیری (در مقایسه با روش نمودار پراکندگی دو متغیره)، محاسبه فاصله مهالانوبیس (Mahalanobis) هر فرد است. آماره فاصله مهالانوبیس یعنی D^2 ، «فاصله» چندمتغیری بین هر فرد و میانگین چندمتغیری گروه را (که کانون نامیده می شود) اندازه گیری می کند. هر فرد با استفاده از توزیع مجذور کای با سطح آلفای دقیق ۰۰۱/. ارزیابی می شود. افرادی را که به این آستانه معنی داری می رسند می توان به عنوان موارد پرت چندمتغیری تلقی کرد و به احتمال باید از نمونه حذف شود(میزر، گامست و گارینو،۱۰۶:۱۳۹۱).

مثال

در تحقیقی سه متغیر اصلی وجود دارد که شامل متغیرهای رضایت شغلی، استرس شغلی و تعهد شغلی است. می خواهیم وجود داده پرت چند متغیره را در متغیرهای فوق بسنجیم.

برای آزمودن وجود داده های پرت چندمتغیری، از فرمان رگرسیون استفاده می کنیم.

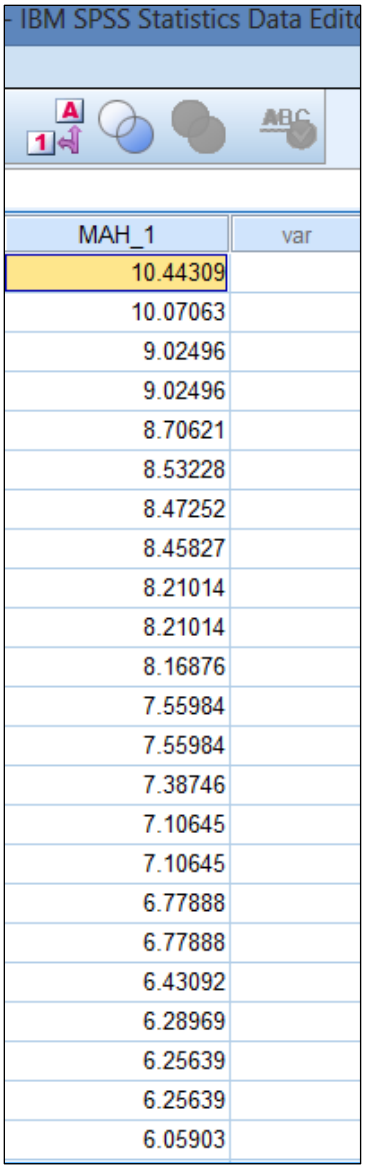
نحوه اجرای دستور محاسبه فاصله مهالانوبیس در SPSS

دستور Analyze ---> Regression ---> Linear را اجرا می کنیم. در کادر ایجاد شده سه متغیر رضایت شغلی، استرس شغلی و تعهد شغلی را وارد کادر Independent می کنیم. یک متغیر را هم وارد کادر Dependent می کنیم. اهمیتی ندارد که چه متغیری را وارد کادر Dependent می کنیم. تعریف یک متغیر وابسته تنها جهت اجرای رگرسیون است و غیر از سه متغیر رضایت شغلی، استرس شغلی و تعهد شغلی می توانیم هر متغیر دیگری را وارد کادر Dependent کنیم.

سپس گزینه Save را انتخاب و در کادر Distance گزینه Mahalanobis را فعال می کنیم. در انتها گزینه Continue و Ok را انتخاب می کنیم.

حاصل این دستورات، خروجی رگرسیون و یک متغیر جدید در فایل داده ها است. متغیر جدید ایجاد شده« Mah_1 » نام دارد که در فایل داده ها تشکیل می شود و آخرین متغیر است.

مقادیر بدست آمده برای فاصله مهالانوبیس با توزیع مجذور خی (کای اسکوئر) مقایسه می شود (جدول توزیع مجذور خی در انتهای برخی کتب آماری وجود دارد). برای این مقایسه ابتدا درجه آزادی فاصله مهالانوبیس را بدست می آوریم که از تفریق تعداد متغیرهای مستقل (وارد شده در کادر Independent رگرسیون) منهای عدد یک بدست می آید. در این مثال ما سه متغیر مستقل داریم و در نتیجه درجه آزادی برابر با ۲ است(۳ متغیر مستقل داشتیم که از عدد ۱ کم می شود و عدد بدست آمده درجه آزادی نام دارد که برابر با عدد ۲ است). مقدار خی دو متناظر با درجه آزادی ۲، عدد است و هر مورد یا پاسخگویی که فاصله مهالانوبیس آن از عدد بیشتر باشد داده پرت محسوب می شود. برای یافتن اعداد بزرگتر از در فایل داده ها و در متغیر Mah_1 بهتر است بر روی نام متغیر در پنجره Data view کلیک راست کنیم و گزینه Sort Descending را انتخاب کنیم تا داده ها از زیاد به کم مرتب شوند. حال کافیست اعداد بزرگ تر از را پیدا کنیم و سپس مورد یا پاسخگوی مورد نظر را از فایل داده ها حذف کنیم. نتایج بدست آمده نشان می دهد که بالاترین عدد بدست آمده برابر با است که از مقدار کمتر بوده و نشان می دهد داده پرت چندمتغیری در داده های ما وجود ندارد.



منبع:

میزر، لاورنس اس و گامست، گلن و گارینو، ا.جی. (۱۳۹۱) پژوهش چندمتغیری کاربردی (طرح و تفسیر)، ترجمه حسن پاشا شریفی و دیگران، تهران: رشد.