دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مدیریت علم و فناوری

گزارش کار هفته سوم

پیش پردازش داده

نگارش

رضا اکبری مقدم

استاد

دکتر مهدی قطعی

آبان ماه ۹۹

فهرست

فهرست اشکال

چکیده

شروع هر نوع کار و عملیاتی در مرحله اول، دارای یک سری مقدمات و پیش‌نیازها است.
داده‌کاوی (Data Mining) نیز از این قانون مستثنی نبوده و نیازمند آماده‌سازی و پردازش‌های مقدماتی است. در علم داده‌کاوی، تمامی داده‌هایی که برای هدف مورد نظر استفاده خواهند شد، باید پیش از شروع پردازش با استفاده از روش‌هایی، آماده و تنظیم و یا به اصطلاح پیش‌پردازش (Preprocess)شوند. مرحله آماده‌سازی داده‌ها قبل از پردازش را، پیش‌پردازش (Preprocessing)می‌گویند. پیش‌پردازش نقشی اساسی در روند پردازش داده‌ها و نتایج حاصل از آن‌ها ایفا می‌کند .

مقدمه

معرفی مجموعه داده New York City Airbnb Open Data

از سال ۲۰۰۸ مهمان ها و میزبان ها از Airbnb برای مسافرت ها و پیدا کردن محل خاص و شخصی خودشان استفاده میکنند.

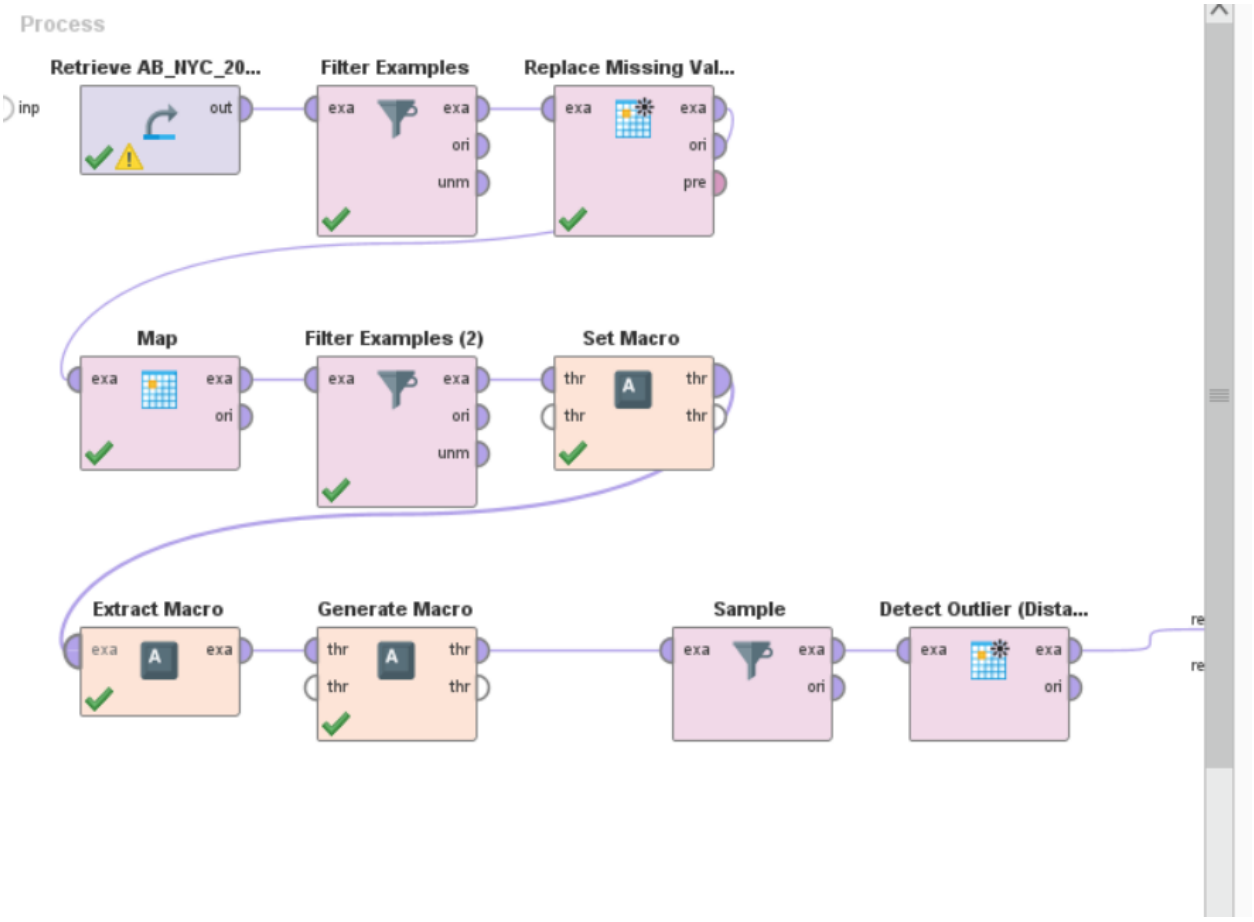
این مجموعه داده توصیفی از فعالیت های شهر نیویورک سیتی در سال ۲۰۱۹ میباشد که شامل ۱۶ ویژگی زیر میباشد :

- Id : شماره شناسایی مهمان
- Name : نام مهمان
- Host_id : شماره شناسایی میزبان
- Host_name : اسم میزبان
- neighbourhood_group : گروه محله
- neighbourhood : محله
- latitude : عرض جغرافیایی
- Longitude : طول جغرافیایی
- room_type : نوع اتاق
- Price : قیمت

- minimum_nights : حداقل تعداد شب های اجاره شده
- number_of_reviews : تعداد مراجعات
- last_review : آخرین مراجعه
- reviews_per_month : مراجعات در هر ماه
- calculated_host_listings_count : تعداد دفعات اجاره دادن میزبان
- availability_365 : دسترسی در ۳۶۵ روز سال

پیش پردازش داده ها

با استفاده از نرم افزار rapidminer داده را وارد میکنیم.



شکل ۱

Data cleaning

اگر به معنای ویژگی ها توجه کنیم متوجه میشویم که برخی از ویژگی ها بازه ی ثابت و مشخصی دارند لذا نباید از این بازه خارج باشند.

در ویژگی های reviews_per_month و minimum_nights به ترتیب بازه های۰-۳۰ و ۰-۳۶۵ را خواهیم داشت که در این ویژگی شاهد برخی outlier هستیم که با استفاده از اپراتور فیلتر مشخص میکنیم که خارج از این بازه را نادیده بگیرد.

در ادامه برخی داده ها از دست رفته میباشند که در شکل ۲ مشخص شده است.

Retrieve AB_NYC_2019.output (output)
Meta data: Data Table
- *Source:* //Local Repository/data/AB_NYC_2019

Number of examples = 48610
16 attributes:
Note: Some of the nominal values in this set were discarded due to performance reasons. You can change this behaviour in the preferences (`rapidminer.general.md_nominal_values_limit`).
Generated by: Retrieve AB_NYC_2019.output ← Retrieve AB_NYC_2019.output
Data: SimpleExampleSet: 48610 examples, 16 regular attributes, no special attributes

Role	Name	Type	Range	Missings	C
	id	integer	=[2539 – 3...	= 15	
	name	polyno...	⊇[* ORIGIN...	= 75	
	host_id	integer	=[2438 – 2...	= 474	
	host_name	polyno...	⊇[Abdul, Ad...	= 495	
	neighbourh...	polyno...	=[Bronx, Br...	= 474	
	neighbourh...	polyno...	⊇[Allerton, ...	= 474	
	latitude	real	=[40.500 – ...	= 474	
	longitude	real	=[-74.244 –...	= 474	
	room_type	polyno...	=[Entire ho...	= 474	
	price	integer	=[0 – 10000]	= 474	
	minimum_...	integer	=[1 – 1250]	= 474	
	number_of...	integer	=[0 – 629]	= 474	
	last_review	date	=[Mar 28, 2...	= 10404	
	reviews_pe...	real	=[0.010 – 5...	= 10404	
	calculated_...	integer	=[1 – 327]	= 474	
	availability_...	integer	=[0 – 365]	= 474	

شکل ۲

با استفاده از اپراتور replace missing به جای داده هایی که خالی یا از دست رفته هستند مقادیر میانگین ویژگی را تعیین میکنیم.

با بررسی داده ها متوجه میشویم برخی از مقادیر ویژگی قیمت صفر میباشند که بی معنی است و هر اتاق بایستی کرایه مشخصی داشته باشد. لذا برای این مقادیر صفر با استفاده از اپراتور map مقدار میانگین قیمت را قرار میدهیم.

	Name	Type	Range	Missings
	price	# integer	≤[0 – 10000]	= 0
	last_review	date	=[Mar 28, 2...	= 104
	id	# integer	=[2539 – 3...	= 0
	name	polyno...	⊇[* ORIGIN...	= 0
	host_id	# integer	=[2438 – 2...	= 0
	host_name	polyno...	⊇[Abdul, Ad...	= 0
	neighbourh...	polyno...	=[Bronx, Br...	= 0
	neighbourh...	polyno...	⊇[Allerton, ...	= 0
	latitude	# real	=[40.500 – ...	= 0
	longitude	# real	=[-74.244 –...	= 0
	room_type	polyno...	=[Entire ho...	= 0
	minimum_...	# integer	=[1 – 1250]	= 0
	number_of...	# integer	=[0 – 629]	= 0
	reviews_pe...	# real	=[0.010 – 5...	= 0
	calculated_...	# integer	=[1 – 327]	= 0
	availability_...	# integer	=[0 – 365]	= 0

شکل ۳

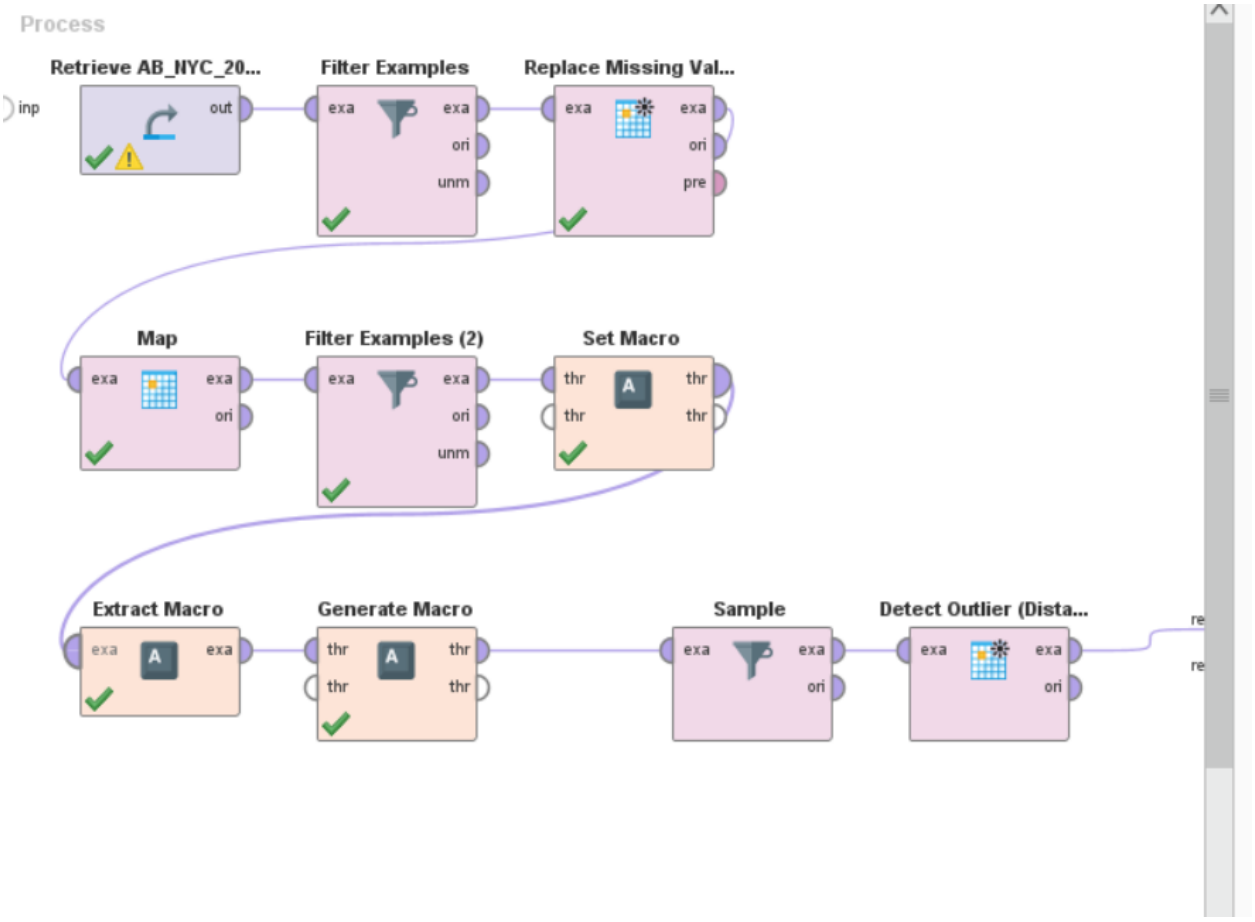
با توجه به شکل ۳ متوجه میشویم که هنوز داده های ویژگی تاریخ به دلیل اینکه نمیتوان میانگینی از آن گرفت از دست رفته میباشند. لذا با توجه به حجم بالای نمونه ها و تعداد از دست رفته های این ویژگی با حذف آن ها مشکل خاصی در نتایج خروجی این تحلیل بوجود نمی آید. لذا با استفاده از اپراتور فیلتر این داده ها را حذف میکنیم.

Role	Name	Type	Range	Missings	C
	price	integer	≤[0 – 10000]	= 0	
	last_review	date	=[Mar 28, 2...	= 0	
	id	integer	=[2539 – 3...	= 0	
	name	polyno...	⊇[* ORIGIN...	= 0	
	host_id	integer	=[2438 – 2...	= 0	
	host_name	polyno...	⊇[Abdul, Ad...	= 0	
	neighbourh...	polyno...	=[Bronx, Br...	= 0	
	neighbourh...	polyno...	⊇[Allerton, ...	= 0	
	latitude	real	=[40.500 – ...	= 0	
	longitude	real	=[-74.244 –...	= 0	
	room_type	polyno...	=[Entire ho...	= 0	
	minimum_...	integer	=[1 – 1250]	= 0	
	number_of...	integer	=[0 – 629]	= 0	
	reviews_pe...	real	=[0.010 – 5...	= 0	
	calculated_...	integer	=[1 – 327]	= 0	
	availability_...	integer	=[0 – 365]	= 0	

شکل ۴

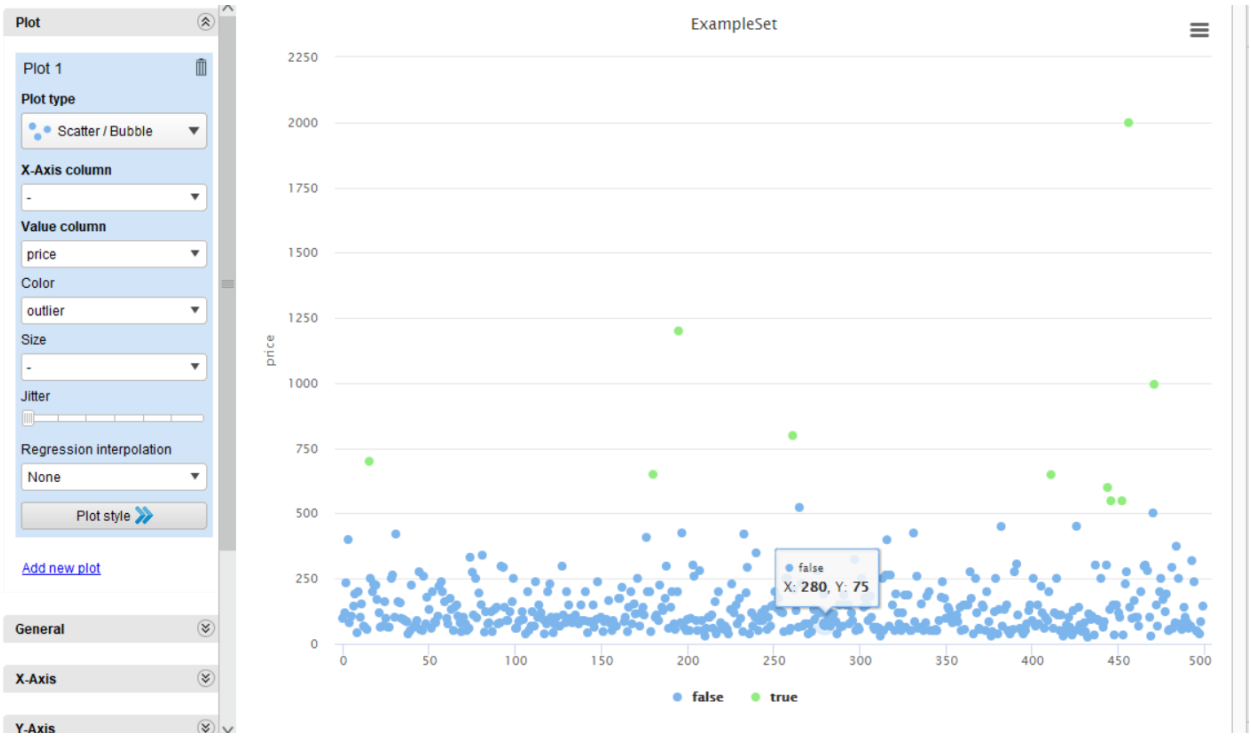
همانطور که در شکل ۴ مشاهده میشود دیگر داده از دست رفته ای در مجموعه داده ما موجود نمیباشد و یک مجموعه داده صحیح در اختیار داریم.

سپس بر اساس ویژگی قیمت بررسی میکنیم که آیا در این ویژگی outlier موجود میباشد یا خیر. در شکل ۵ توسط اپراتور detect outlier یک ویژگی جدید به اسم outlier ایجاد میشود که مقادیر true و false دارد که به معنای این میباشد یک ویژگی مقدار outlier دارد یا خیر.



شکل ۵

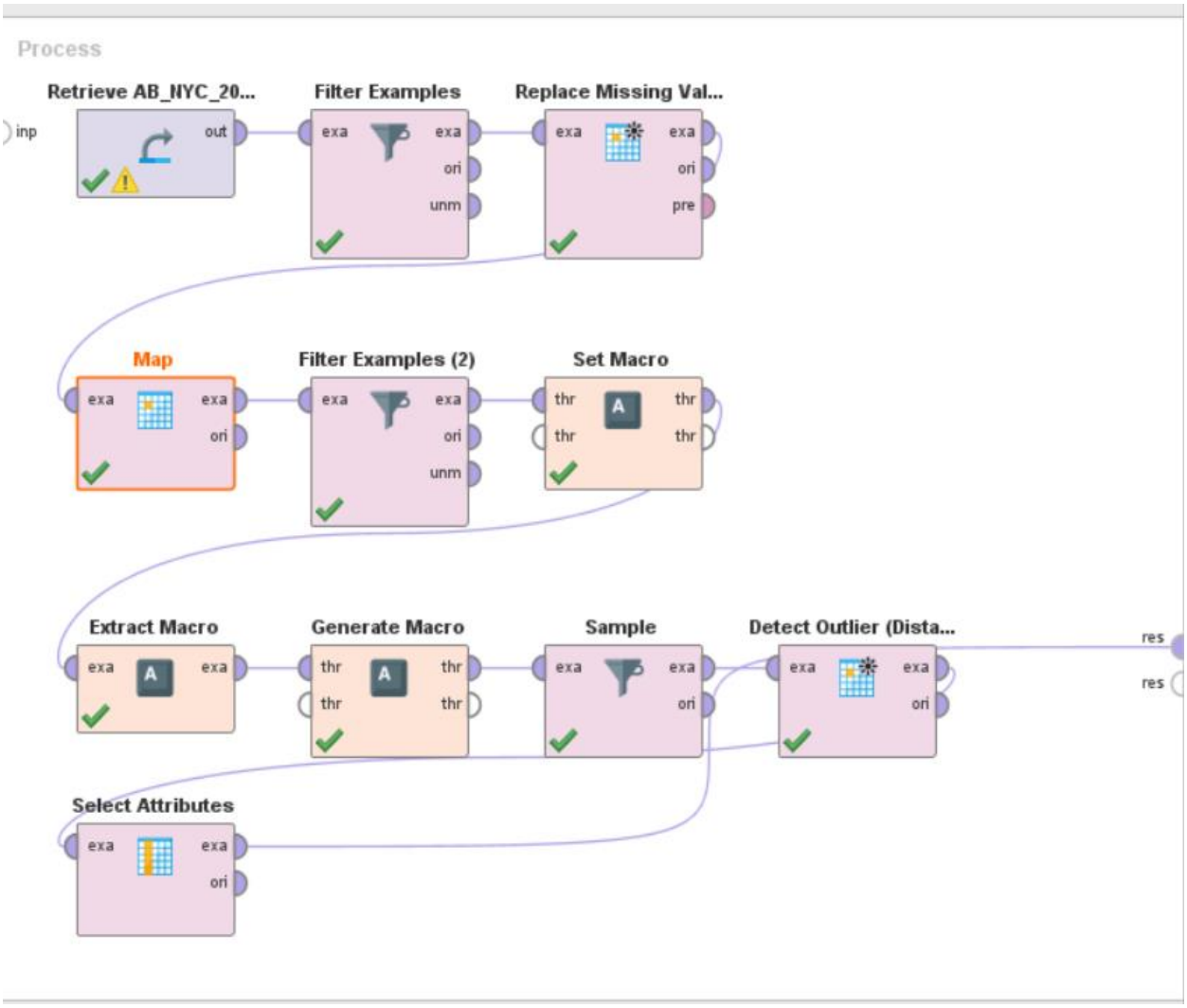
همانطور که در شکل ۶ مشاهده میشود در نمودار scatter این ویژگی مقادیر outlier با مقدار true مشخص شده اند که این مقادیر بالاتر از ۶۰۰ میباشند. لذا با استفاده از اپراتور select attribute مقادیر بیشتر از ۶۰۰ قیمت را حذف میکنیم.



شکل ۶

Data reduction

در این قسمت میخواهیم با استفاده از تکنیک های data reduction ابعاد مجموعه داده را کاهش دهیم.



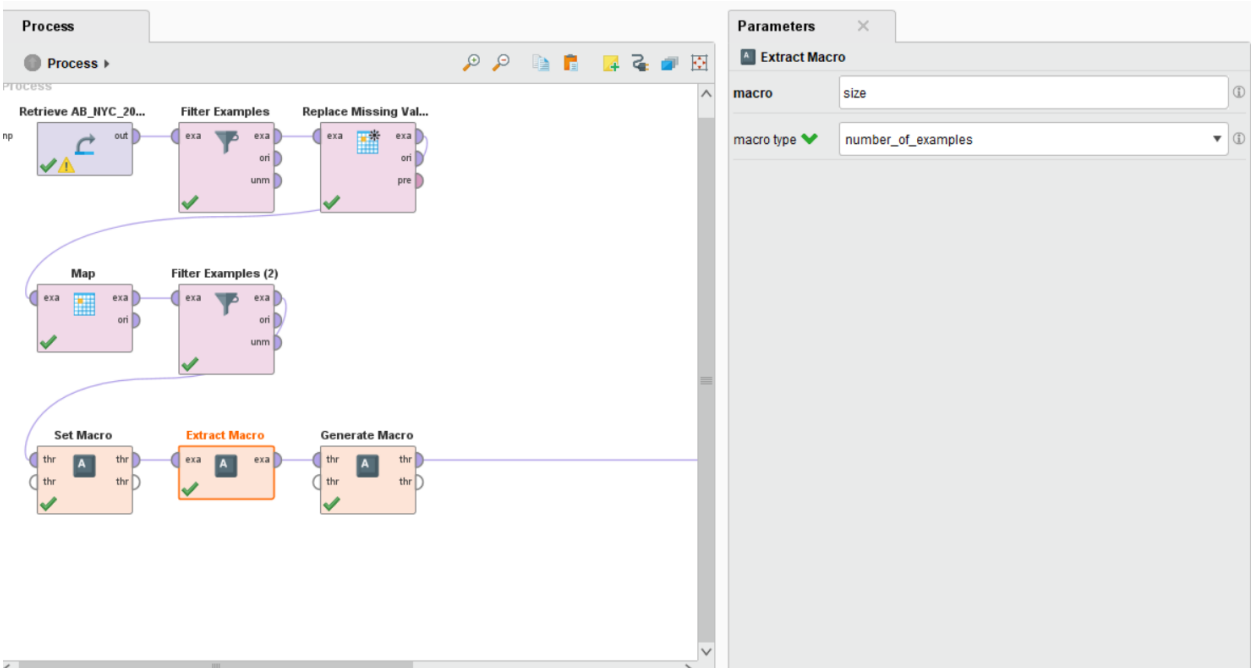
شکل ۷

در این مجموعه داده برخی ویژگی ها کارایی چندانی ندارند مانند طول و عرض جغرافیایی، نام میزبان و نام. چرا که در برخی ویژگی های دیگه مانند محله مکان نمونه مشخص شده است لذا با استفاده از اپراتور select این ویژگی ها را از مجموعه داده نادیده میگیریم.

Data transformation

در این قسمت میخواهیم نمونه های موجود در این مجموعه داده را کاهش دهیم تا تحلیل آسانتر شود.

با استفاده از اپراتورهای ماکرو که در شکل ۵ مشخص شده است میتوانیم به صورت پویا نمونه ها و ویژگی های مجموعه داده را کاهش دهیم. در این بخش ما با استفاده از ماکرو نمونه های این مجموعه را به ۵۰ درصد کاهش میدهیم و با استفاده از اپراتور sample این تغییرات را اعمال میکنیم.



شکل ۸

در نهایت یک مجموعه داده پیش پردازش شده در اختیار داریم تا به صورت کارآمد و بهینه از داده ها استفاده شود.