Suicide Rates Overview 1985 to 2016

ارائه دهنده: پگاه خزائی

استاد: دکترحسین حاجی ابوالحسن

راهنما: آقای عرفان رستمی

گزارش کار- نظریه ی یادگیری ماشین

دانشگاه شهید بهشتی - دانشکده ی ریاضی وعلوم کامپیوتر

تابستان ۹۹

مقدمه

مجموعه داده ای که روی آن کار کردیم، مربوط به خودکشی افراد در طی سال های ۱۹۹۵ تا ۲۰۱۶ است.

مجموعه داده ۱۲ ویژگی دارد که عبارتند از:

- کشور
- سال
- جنسیت
- گروه سنی
- تعداد خود کشی ها
- جمعیت
- نرخ خودکشی: تعداد خودکشی ها در هر ۱۰۰۰۰۰ نفر
- کشور- سال: کدی که حاوی نام کشور به علاوه سال خودکشی است.
- شاخص "Gross Domestic Product" یا به طور مختصر "GDP". به معنای تولید ناخالص داخلی.
(ارزش پولی تمام شده کالاها و خدمات تولید شده در داخل مرزهای یک کشور در یک بازه زمانی مشخص است.)
- GDP per capita: سرانه تولید ناخالص داخلی. برای محاسبه این شاخص کافی است که GDP کل را به جمعیت آن کشور تقسیم کرد.
- HDI برای سال: شاخص توسعه انسانی (Human Development Index) یک شاخص مرکب آماری از شاخص های امید به زندگی، تحصیلات و درآمد سرانه است.
- نسل: نام نسل مربوطه

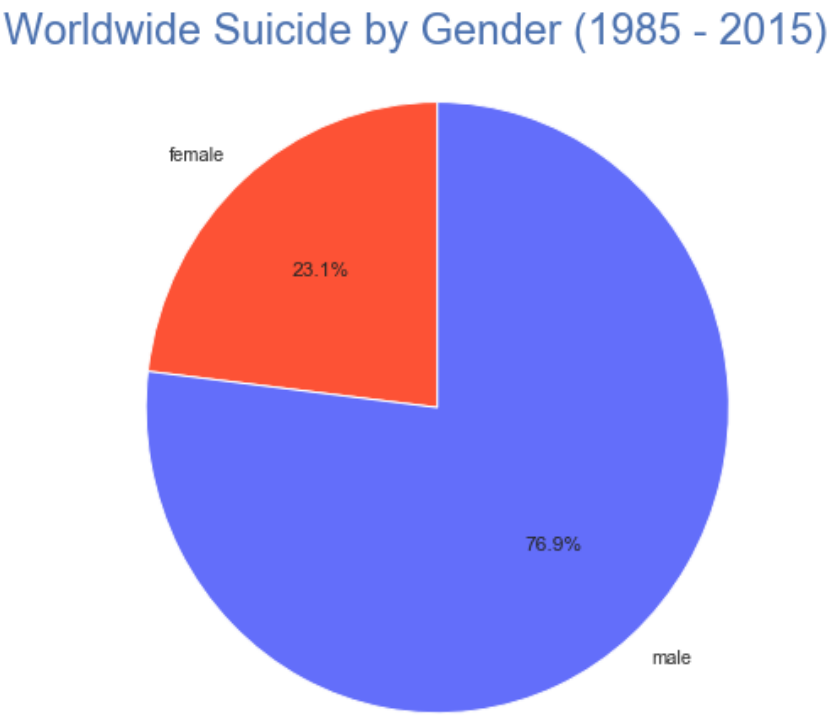
EDA (Exploratory Data Analysis)

در طول این تحلیل ابتدا داده ها را تمییز میکنیم و تحول لازم را ایجاد میکنیم. (ما هر زمان که به تجزیه و تحلیل لازم باشد داده های خود را تغییر خواهیم داد.)
ما تجزیه و تحلیل داده های اکتشافی را انجام خواهیم داد و سعی خواهیم کرد چندین بینش مفید از مجموعه داده ها را استخراج کنیم.

ستون مربوط به ویژگی "کشور- سال" تکراری است. به دلیل وجود ویژگی کشور و سال به طور جداگانه در مجموعه ی داده آن را حذف کردیم.
ستون مربوط به HDI با ۱۹۴۵۶ مقدارازدست رفته تنها ویژگی در مجموعه داده است که مقادیراز دست رفته (Missing Values) دارد.
میتوانستیم ستون نظیر HDI را از مجموعه داده ی خود حذف کنیم.
همچنین میتوانستیم آن ها را با یک مقدار پر کنیم. برای مثال با مقدار صفر یا با میانگین سایر مقادیر.
که ما در تحلیل خود مقادیر از دست رفته را با میانگین مقادیر ستون HDI پر کردیم.
با بررسی بیشتر از مجموعه داده ها به دست می آید که:

- ۱۰۱ "کشور" یکتا (Unique) در مجموعه داده وجود دارد.
- میزان خودکشی در آقایان از خانوم ها بیشتر است.
- ویژگی "گروه سنی" دارای ۶ مورد یکتا است.
- در ستون مربوط به ویژگی نسل، ۶ مورد از نسل ها را مشاهده میکنیم.

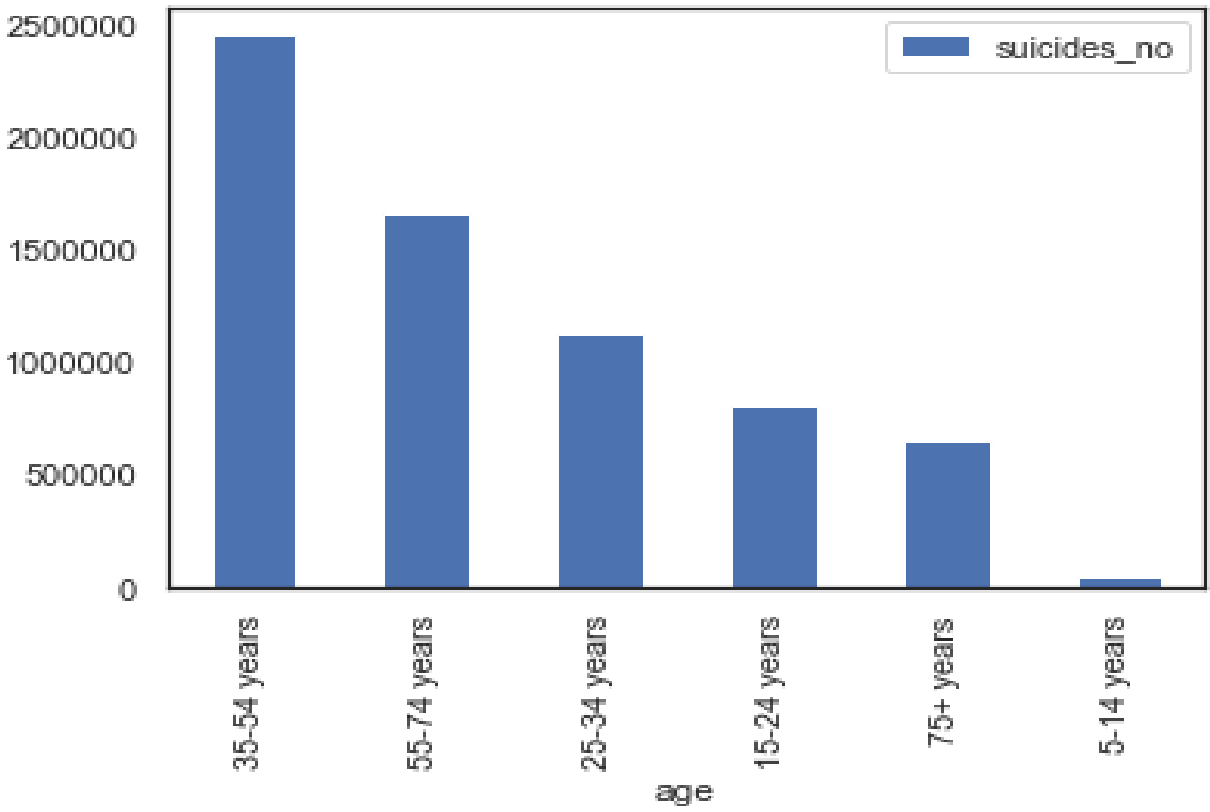
نمودار دایره ای به دست آمده از بررسی مجموعه ی داده، گویا ی این است که ۷۶/۹ درصد از افرادی که خودکشی میکنند را آقایان تشکیل میدهند و ۲۳/۱ درصد این افراد شامل خانم ها هستند.



شکل۱. نمودار دایره ای میزان خودکشی افراد بر اساس جنسیت

خودکشی در گروه های سنی مختلف به چه شکل است؟

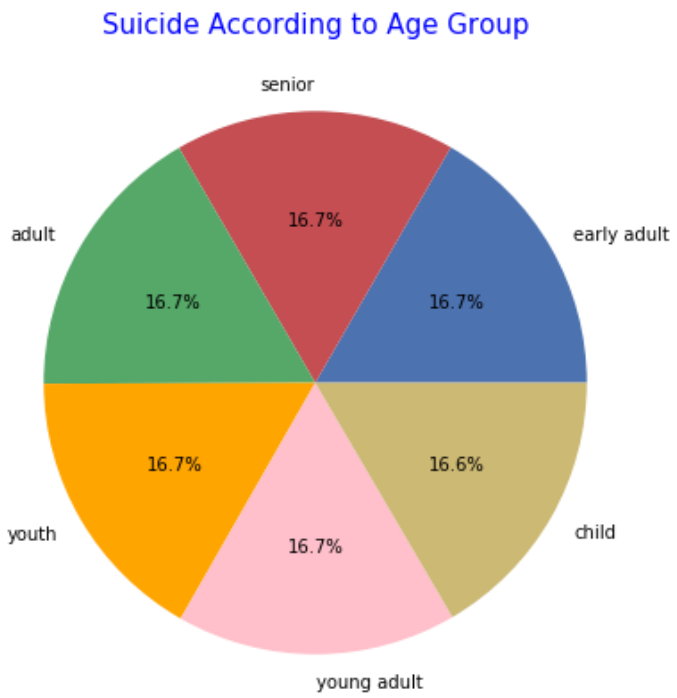
نتیجه را در نمودار میله ای که در شکل آمده است می بینیم.



شکل۲. نمودار میله ای میزان خودکشی گروه های سنی مختلف

همانطور که مشاهده میکنید، بیشترین میزان خودکشی در میان گروه سنی ۳۵ تا ۵۴ سال و بعد از آن در بین افراد با گروه سنی ۵۵ تا ۷۴ سال است.

میتوانیم گروه های سنی مختلف را نام گذاری کنیم و برای نمایش بهتر، آن را با نمودار دایره ای نیز نمایش دهیم.



شکل۳. نمودردایره ای درصد خودکشی گروه های سنی مختلف

۵-۱۴ سال (child)

۱۵-۲۴ سال (youth)

۲۵-۳۴ سال (young adult)

۳۵-۵۴ سال (early adult)

۵۵-۷۴ سال (adult)

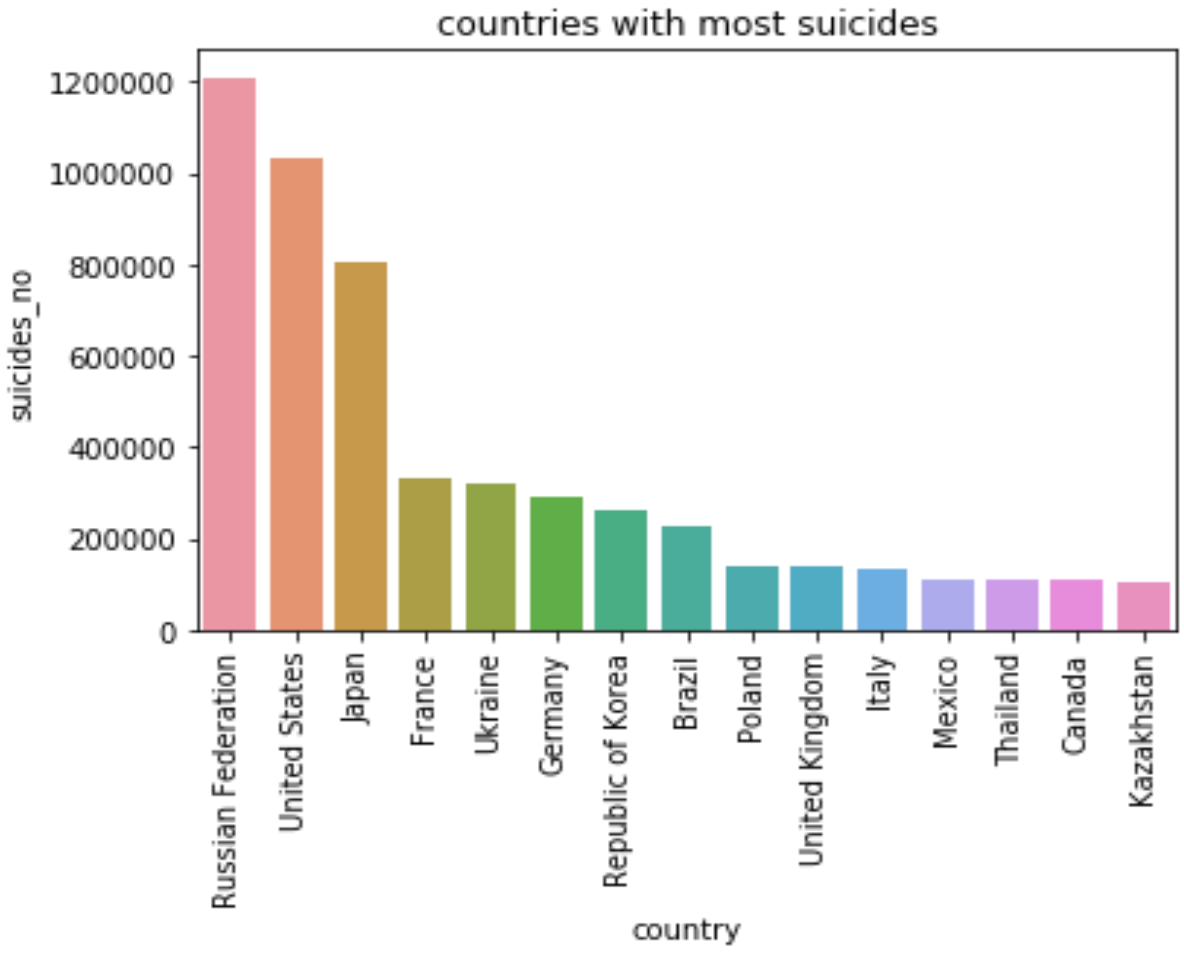
۷۵+ سال (senior)

با برسی کلی تر به این نتیجه میرسیم که تعداد خود کشی ها در بین گروه سنی ۳۵-۵۴ با بیشترین میزان خودکشی از سال ۱۹۹۵ تا سال ۲۰۱۶ برابر است با مقدار ۲۴۵۲۱۴۱.

و این میزان در بین افراد ۵-۱۴ سال که کمترین میزان خودکشی را در این سال ها داشته اند برابر با مقدار ۵۲۲۶۴ است.

۱۵ کشور با بیشترین میزان خودکشی کدام کشور ها هستند؟

با کشیدن نمودار میله ای و بررسی ۱۵ کشور با بیشترین تعداد خودکشی اینگونه به دست می آید که:

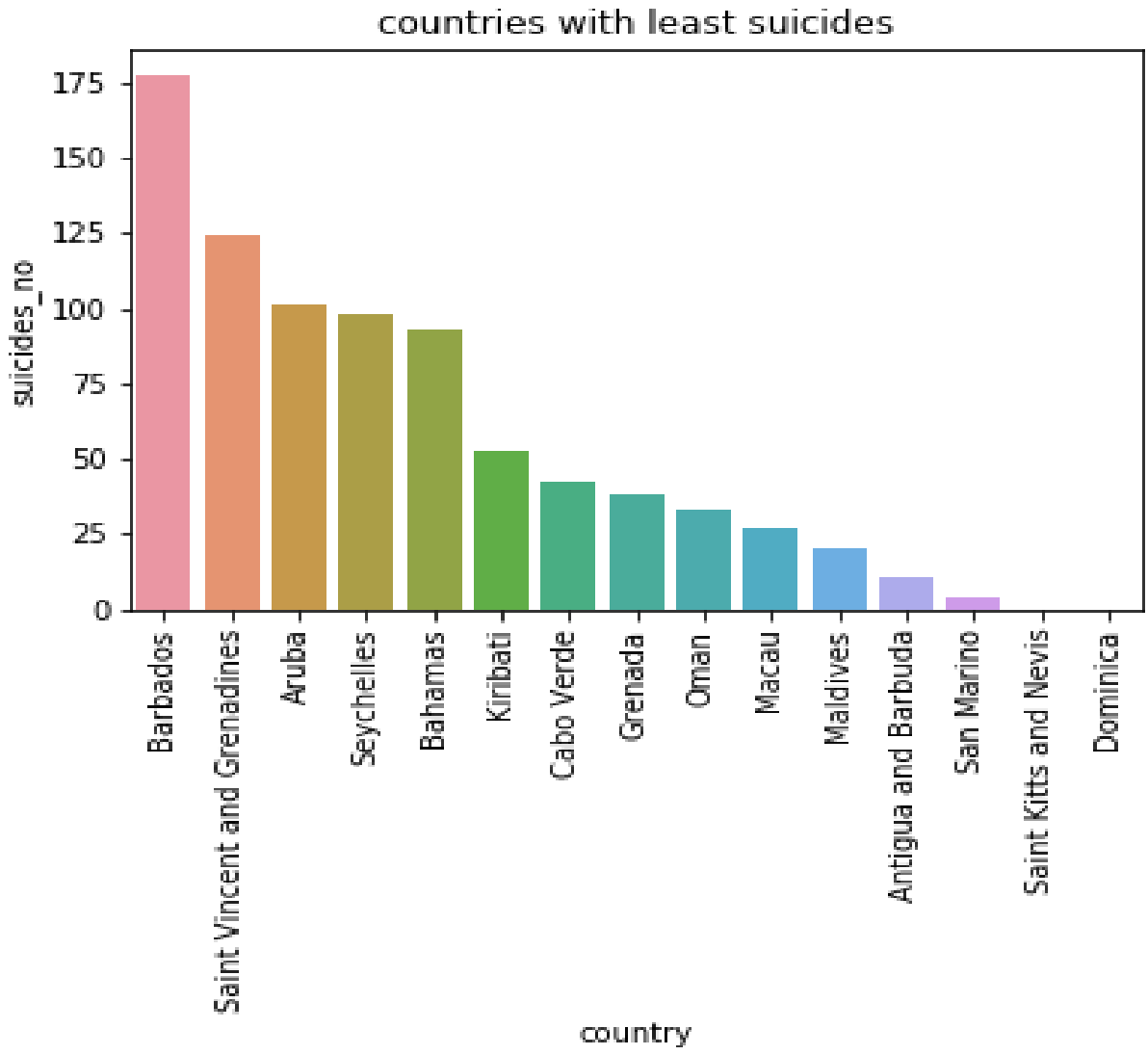


شکل ۴. نمودار میله ای میزان خودکشی در کشورهای مختلف

میزان خودکشی در سه کشور روسیه، ایالات متحده و ژاپن از سایر کشورها بیشتر بوده است.

۱۵ کشور با کمترین میزان خودکشی کدام کشور ها هستند؟

با کشیدن نمودار میله ای و بررسی ۱۵ کشور با کمترین تعداد خودکشی اینگونه به دست می آید که:

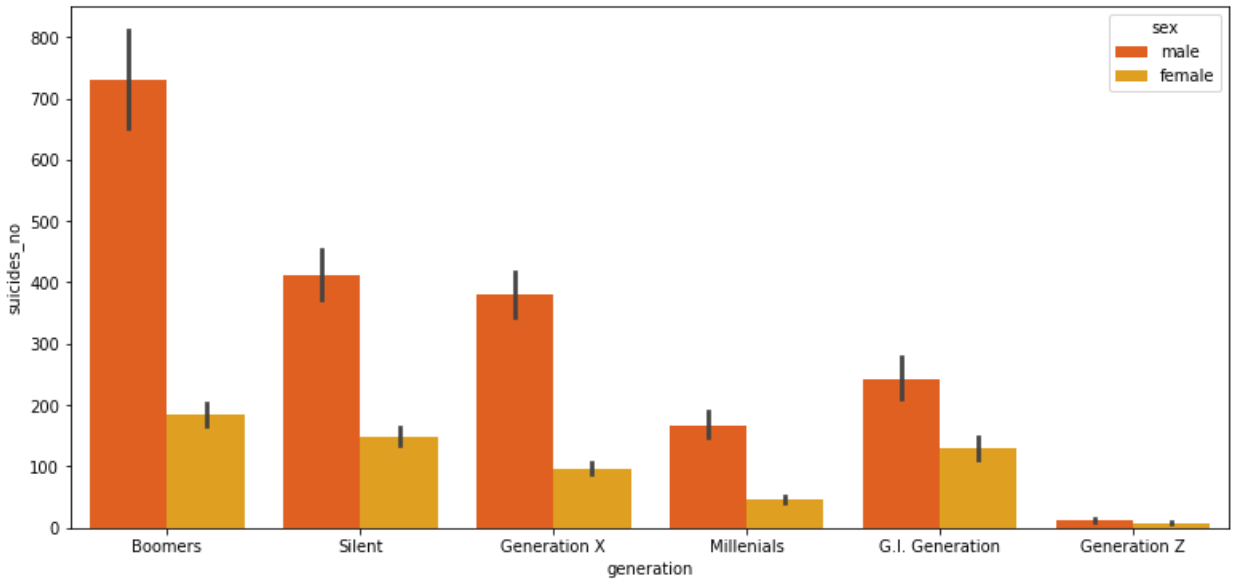


شکل ۵. نمودار میله ای میزان خودکشی در کشورهای مختلف

در دو کشور **Dominica** و **Saint Kitts and Nevis** تعداد خودکشی ها صفر بوده است!

یکی از ستون ها ی مجموعه داده ی ما مربوط به ویژگی نسل است.

از مقایسه میزان خودکشی در بین نسل های مختلف این چنین به دست می آید که:



شکل ۶. نمودار میله ای میزان خود کشی افراد نسل های مختلف

G.I Generation : متولد سالهای ۱۹۲۴-۱۹۰۱، آنها کسانی هستند که در بزرگسالی افسردگی بزرگ و جنگ جهانی دوم را تجربه کردند.

Silent Generation : در بین سالهای ۱۹۴۵-۱۹۲۴، در دوران خوشبختی پس از جنگ به دنیا آمدند. کودکان نسل سکوت در شرایط پیچیده جنگ و رکود اقتصادی بزرگ شدند.

Boomers : آنهایی که بعد از جنگ جهانی دوم به دنیا آمدند و به هیپی گری روی میاوردند.

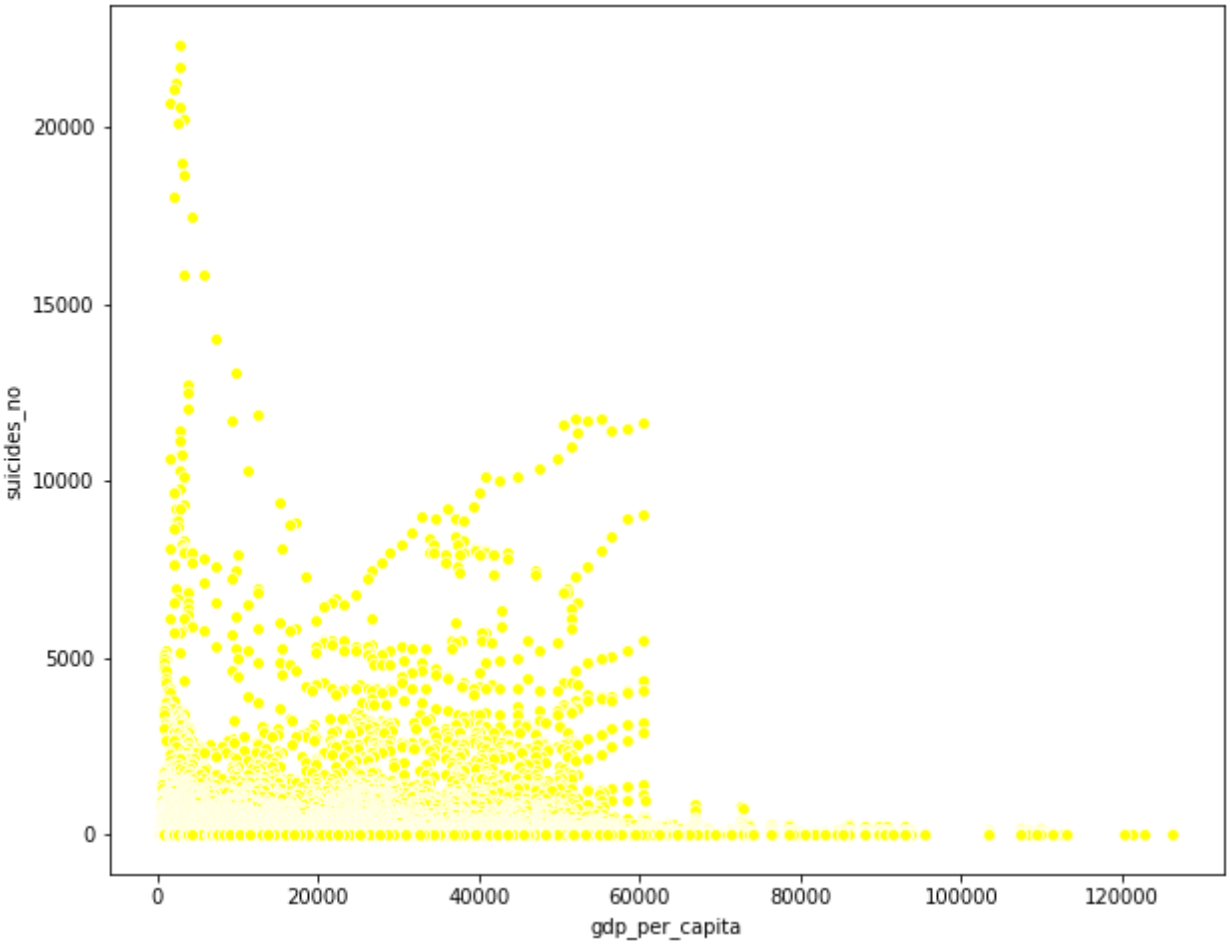
Generation X : در بین سالهای ۱۹۶۵ و ۱۹۸۰ متولد شدند، آنها بچه های باهوشی بودند که در خیابان، منزوی و اغلب نزد والدین مطلقه یا والدینی که تنها روی شغلشان تمرکز داشتند پرورش یافته بودند.

Millennials : محققان و مفسران سالهای تولد را از اوایل دهه ۱۹۸۰ تا اوایل ۲۰۰۰ استفاده می کنند. آنها با فن آوری های پیشرفته آشنا بوده و مصون از آسیب های سنتی میباشند.

Generation Z : این نسلی است که پس از ۱۹۹۵ به دنیا آمد و آنها هرگز دنیایی بدون رایانه و تلفن های همراه نمی شناسند.

همانطورکه مشاهده میکنید میزان خودکشی در نسل جدید کمتر از سایر نسل ها است و درنسل Boomers از سایر نسل ها بیشتر است.

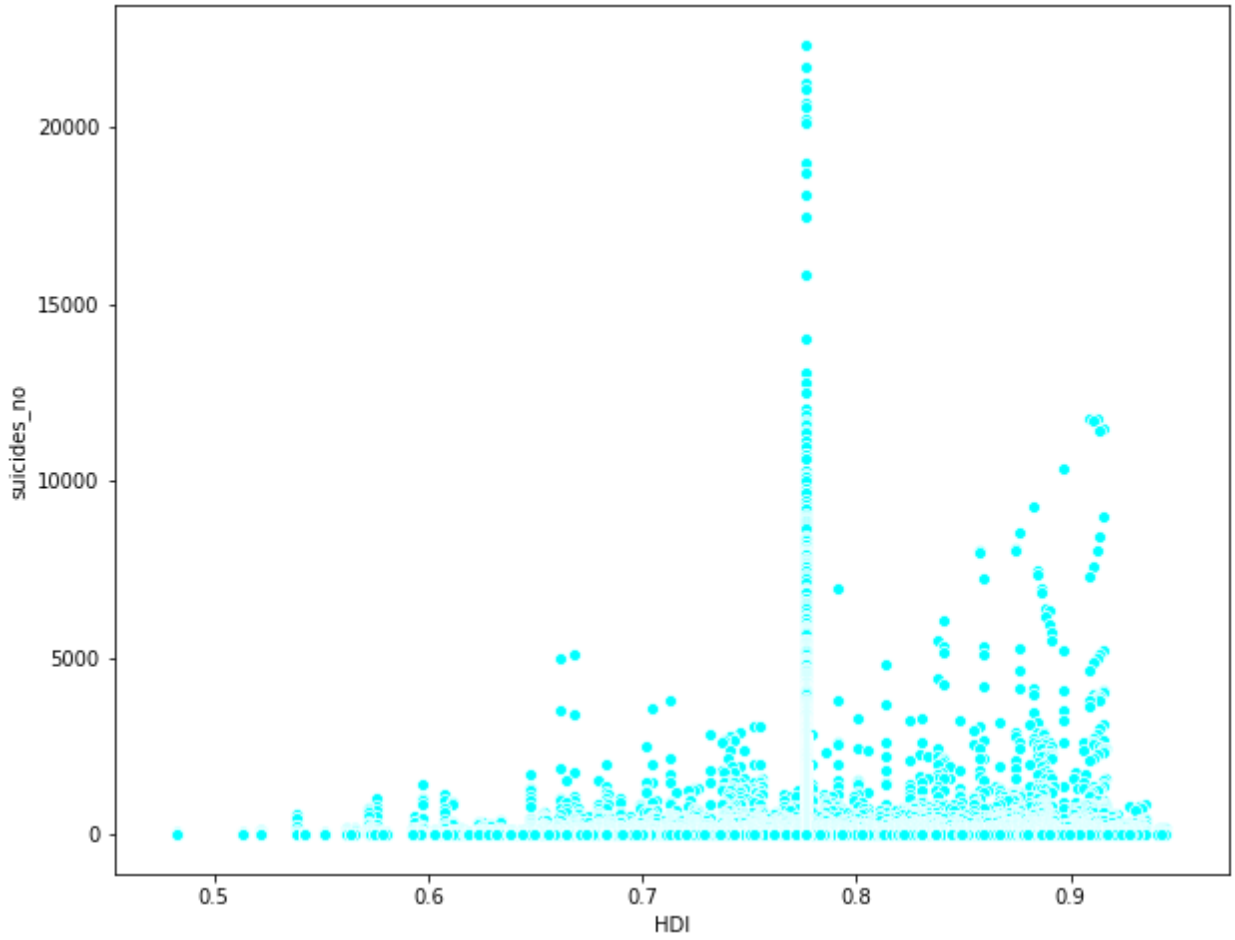
بررسی رابطه ی ببین تعداد خودکشی ها و شاخص GDP.



شکل ۷. تغییرات تعداد خودکشی بر حسب تولید ناخالص داخلی

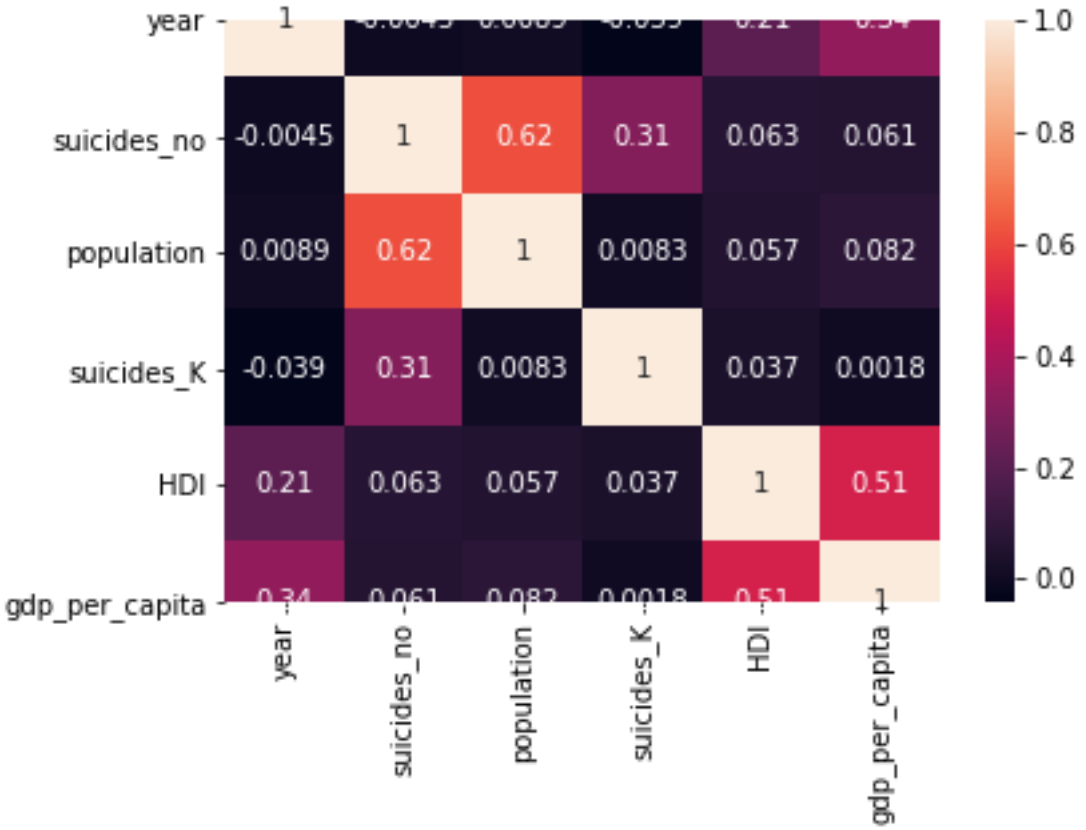
رابطه **"gdp for year"** و **"suicides number"** خطی نیست. از این رو، تولید ناخالص داخلی چیزی نیست که تأثیر واقعی بر میزان خودکشی داشته باشد.

بررسی رابطه ی ببین تعداد خودکشی ها و شاخص HDI.



شکل ۸. تغییرات تعداد خودکشی بر حسب HDI

همبستگی بین ویژگی ها



شکل ۹. ماتریس همبستگی ویژگی ها

همانطور که در ماتریس همبستگی بالا نشان داده شده است، نرخ خودکشی ها و جمعیت همبستگی بسیار ضعیف و نزدیک به صفر را نشان می دهند. (درایه ی سطر سوم و ستون چهارم).

تعداد خود کشی ها و جمعیت، ضریب همبستگی ای برابر با ۰/۶۲ را دارند که نسبتا مقدار قابل توجهی را نشان میدهند. (درایه ی سطر سوم و ستون دوم).

سرانه تولید ناخالص داخلی و شاخص توسعه انسانی ضریب همبستگی ای برابر با ۰/۵۱ را نشان میدهند. (درایه ی سطر ششم و ستون پنجم).

Linear Regression

در این بخش ما سعی خواهیم کرد که مدل های مختلف را با داده های خود متناسب کنیم.

ابتدا می خواهیم مدل رگرسیون خطی را امتحان کنیم.

ما از کتابخانه **SkLearn** هم برای الگوریتم ها و هم برای پردازش داده ها استفاده خواهیم کرد.

برای رگرسیون خطی، ویژگی های:

کشور، جنسیت، جمعیت، سن، **GDP per capita** و نسل را انتخاب میکنیم.

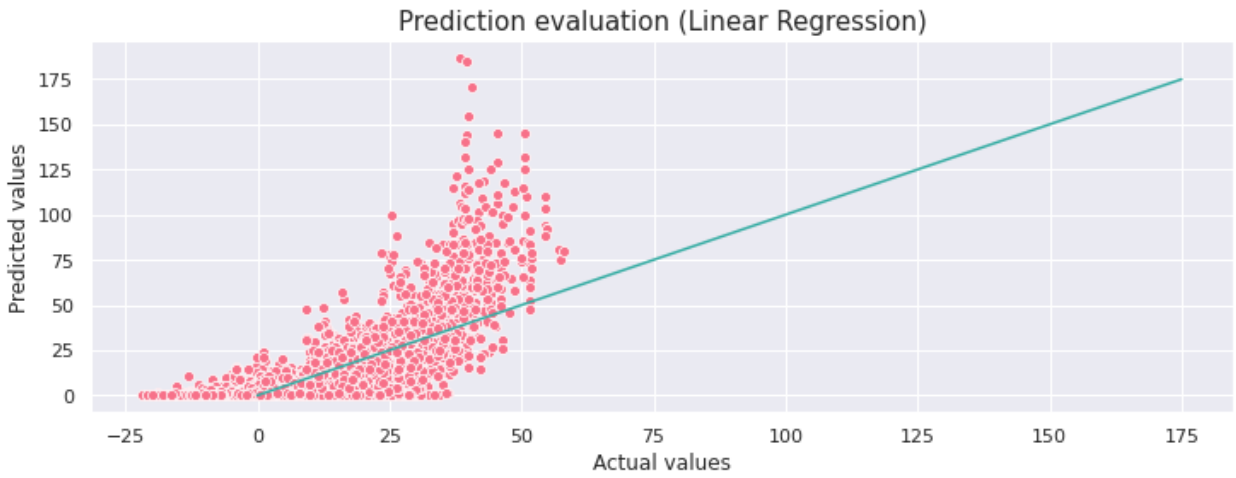
و نرخ خودکشی را در متغیر **y**، ذخیره میکنیم.

همانطور که می دانیم، الگوریتم رگرسیون خطی با ویژگی های **categorical** کار نمی کند.

برای مقابله با این مشکل، ما باید داده های دسته بندی شده خود را به متغیرهای ساختگی تبدیل کنیم.

برای این کار از روش **Pandas get_dummies** استفاده خواهیم کرد.

سرانجام، ما داده های خود را به دو مجموعه تقسیم خواهیم کرد: آموزش و آزمایش. اندازه ها برای داده های آموزش ۸۰٪ و برای تست داده ها ۲۰٪ خواهد بود.

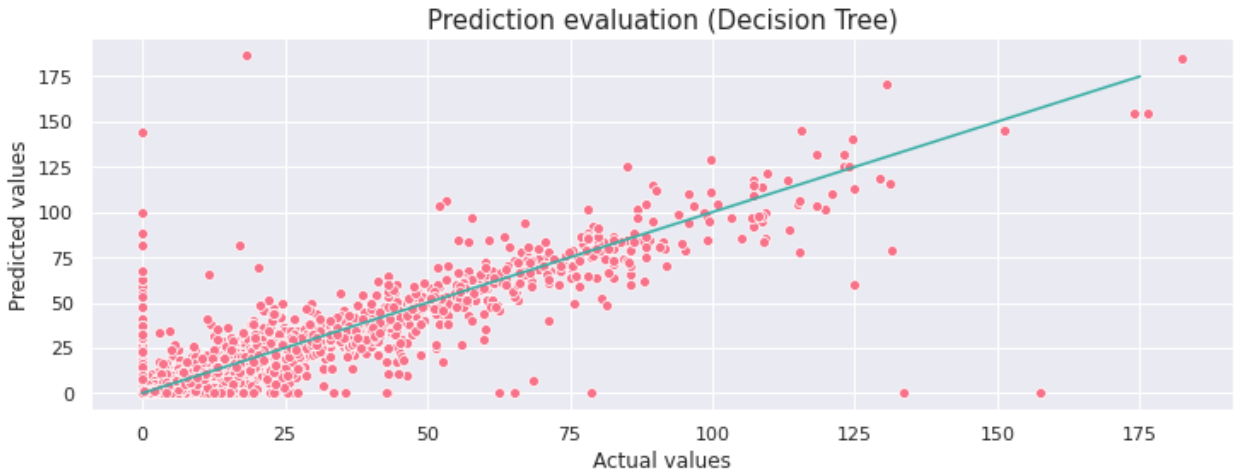


شکل ۱۰. فیت کردن مدل رگرسیون خطی روی داده ها

به نظر نمی رسد که مدل ما کار خوبی را انجام دهد، این ممکن است به این دلیل باشد که ویژگی هایی که انتخاب کردیم به اندازه کافی خوب نیستند یا ممکن است به این دلیل باشد که داده هایی که ما داریم خطی نیستند بنابراین این مدل به قدر کافی مناسب نیست. شاید با درخت تصمیم بهتر شود.

Decision Tree Regressor

ما دقیقا از همان گام هایی که برای رگرسیون خطی انجام دادیم استفاده خواهیم کرد.
سرانجام با ترسیم همان شکل، پیش بینی های خود را با واقعیت مقایسه خواهیم کرد.



شکل ۱۱. فیت کردن مدل درخت تصمیم روی داده ها

پیش بینی ها به وضوح بسیار بهتر از نسخه ای است که رگرسیون خطی تولید می کند. هنوز هم نمیتوانیم بگوییم که مدل نتایج خوبی داشته است. به منظور ایجاد نتایج بهتر، درخت تصمیم گیری نیاز به تنظیم بیشتر دارد. برای تولید نتایج دقیق تر ممکن است داده های ما نیز به تحول بیشتری نیاز داشته باشند یا ممکن است به ویژگی های بیشتری نیاز داشته باشیم.

Decision Tree Classifier

برای کلاس بندی داده ها استفاده کرده کردیم. بعد از انتخاب هایی که برای **X** و **y** داشتیم و گاهاً نتایج ضعیفی که در دقت خروجی مشاهده کردیم، انتخاب ما این بود که:

یک ستون به نام "نرخ مرگ و میر" به مجموعه داده اضافه کنیم.

این ستون این گونه پر میشود که به تک تک اعضای ستون "نرخ خودکشی" نگاه میکنیم و هر جا مقدارش از "میانگین نرخ خودکشی ها" بیشتر بود "۱" را در ستون نرخ مرگ و میر قرار میدهیم و هر جا مقدارش کمتر بود "۰" را قرار میدهیم و اینگونه ستون مربوط به ویژگی نرخ مرگ و میر با صفر و یک پر میشود.

مرحله ی بعدی انتخاب **X** و **y** است.

برای انتخاب **X**، از مجموعه داده ی خود دو ویژگی "نرخ خودکشی" و "نرخ مرگ و میر" را حذف میکنیم. آن را تبدیل به آرایه کرده و در **X** قرار میدهیم.

برای انتخاب **y**، ستون مربوط به "نرخ مرگ و میر" را هدف قرار داده و آن را تبدیل به آرایه کرده و در متغیر **y** قرار میدهیم.

با **fit** کردن مدل به نتیجه ی بسیار خوبی خواهیم رسید.

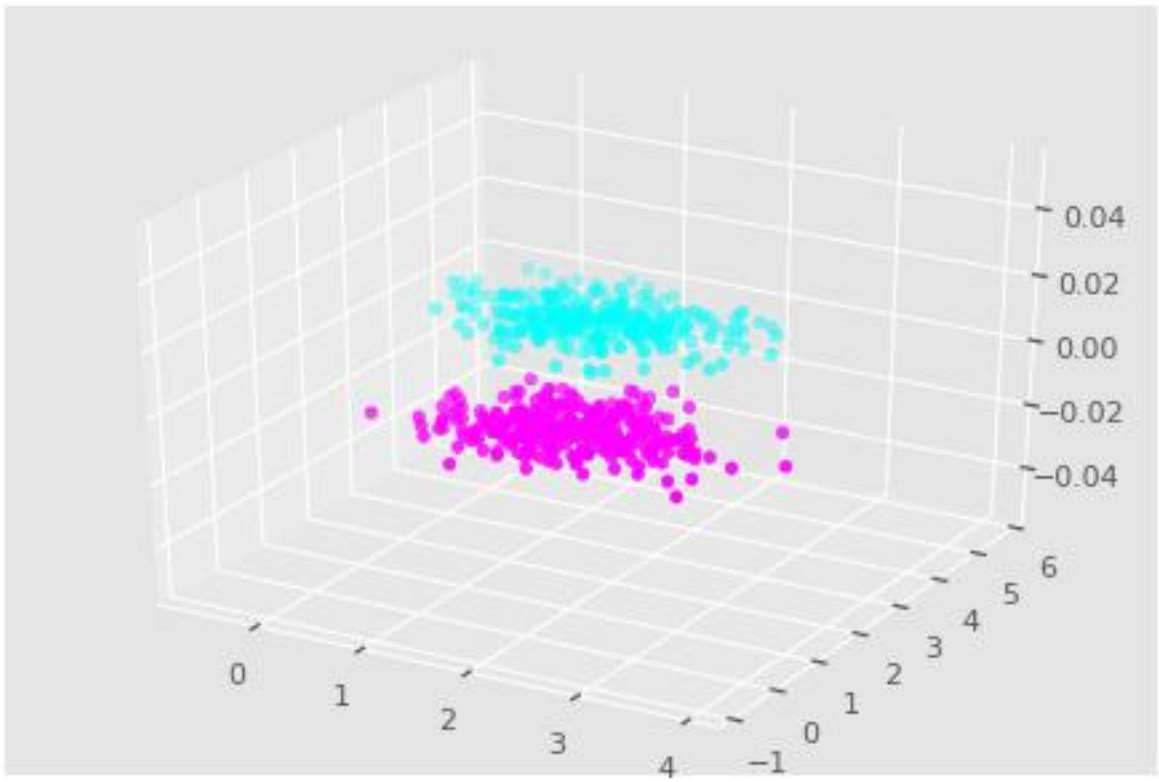
و آن دقت ۹۰ درصد است.

Random Forest Classifier

با انتخاب همین **X** و همین **y** این مدل را روی دیتای خود فیت میکنیم و دقت ۹۷ درصد به دست می آید.

K-means

وظیفه این است که کشورها را به دو گروه طبقه بندی کنیم. گروههایی که تعداد خودکشی بالایی دارند و گروههایی که تعداد خودکشی های کمی دارند. برای این کار باید ستون **"suicides_no"** را از مجموعه داده رها کنیم و آن را بدون برچسب بگذاریم.



شکل ۱۲. نمودار سه بعدی. خوشه بندی داده ها به روش **K-means**

این مدل قادر بود به طور صحیح با ۷۱٪ خوشه بندی کند.

نتیجه گیری

تمیز کردن داده ها (**data cleaning**) بسیار مهم است. تصویر سازی (**Visualizing**) نیز یک قدم بسیار مهم است زیرا درک بسیاری از افراد را برای داده ها آسان تر میکند.

الگوریتم خوشه بندی **K-means** در این مورد آسان بود، زیرا ما دانش حوزه ای داشتیم که تعداد خودکشی های افراد در کشورهای مختلف را به ما می گفت، بنابراین ما مجبور نبودیم تعداد خوشه ها را از قبل مشخص کنیم. با این حال، این حالت همیشه اتفاق نمی افتد. درمورد خودکشی ها و عواملی که بر آنها تأثیر می گذارد، می توان گفت که سن و جنس می تواند برخی از این عوامل باشد اما **HDI** و **GDP** جز این عوامل نیست. زیرا حتی در کشورهایی که **GDP** و **HDI** بالایی دارند، افراد زیادی خودکشی می کنند.

به غیر از این، داده های کافی برای تجزیه و تحلیل بهتر در دسترس نیست ، زیرا عوامل بیولوژیکی، روانی و اجتماعی دیگری نیز وجود دارند که ممکن است باعث خودکشی شوند (نژاد، قومیت، انزوای اجتماعی، مسری، دین و غیره) و همچنین جغرافیایی (اقلیمی).