خطای داده

اشتباه‌های رایج هنگام کار کردن با آمار
راهنمای تشخیص خطا و راه‌های مقابله

داده‌های باز ایران
IRANOPENDATA

خطای داده چیست؟

ما حین کار کردن با داده‌ها در معرض انواع خطا هستیم. نوع ساده خطا، درست بودن یا نبودن خود داده‌ها است. به دلایل مختلفی ممکن است داده‌هایی که در اختیار ما قرار می‌گیرد، عمدا یا سهوا اشتباه باشند یا حتی دستکاری شده باشند. یعنی ما در معرض سوء اطلاعات قرار گرفته باشیم.
به همان نسبت که خطای داده انواع گوناگونی دارد، راه‌های متنوعی هم برای مقابله یا کاهش ریسک آنها وجود دارد، اما پیش از هر چیز باید با انواع رایج و متداول خطاهایی آشنا شد که هنگام کار با داده پیش روی ما قرار دارند.

سهل‌انگاری، خطا، تقلب
خطاهایی که حین پیدا کردن داده‌ها با آنها روبه‌رو می‌شویم. این خطاها در مرحله اول، در زمان پیدا کردن و جمع‌آوری داده‌ها، معمولا به دلیل سهل‌انگاری ما در مراجعه به منابع اصلی یا اشتباه‌های سهوی یا عمدی منابع تولید داده پیش می‌آیند.

اشتباه‌های محاسباتی
اشتباه‌های محاسباتی در زمان کار کردن با داده‌ها و تحلیل و پردازش آنها، یکی از رایج‌ترین انواع خطایی است که ما در معرض آن قرار داریم. این اشتباه‌ها می‌توانند ناشی از اشتباه در عملیات ساده ضرب و تقسیم یا تحت تاثیر استفاده نابجا از ابزارهای آماری به وجود بیاید.

خطای پردازش و ارائه
آخرین مرحله در کار روزنامه‌نگاری داده‌محور اشتباه در انتخاب وسیله مناسب برای ارائه نتیجه کار است. گاهی استفاده از یک نمودار اشتباه یا یک مقایسه غلط می‌تواند، بر خلاف هدف اولیه ما، باعث انتشار اطلاعات نادرست و گمراه‌کننده شود.

داده‌های غلط

۱- خطای منبع؛ از اشتباه‌های سهوی تا دستکاری و جعل آمار

همواره خطر آن وجود دارد که ما در ابتدای کار با داده‌های غلط روبه‌رو شویم. منابع گوناگون ممکن است به دلایل مختلف، از **اشتباه سهوی** تا **دستکاری و جعل آمار و ارقام** داده‌های اشتباه در اختیار ما بگذارند.

سال	۱۳۹۶		۱۳۹۷		۱۳۹۸		۱۳۹۹		۱۴۰۰	
استان	تعداد	نسبت به کل	تعداد	نسبت به کل	تعداد	نسبت به کل	تعداد	نسبت به کل	تعداد	نسبت به کل
کل کشور	۶,۸۰۱,۶۸۴	۱۰۰.۰۰%	۷,۰۶۷,۸۲۷	۱۰۰.۰۰%	۶,۱۶۶,۷۲۹	۱۰۰.۰۰%	۲,۴۱۶,۰۳۱	۱۰۰.۰۰%	۳,۰۴۹,۹۹۹	۱۰۰.۰۰%
آذربایجان‌شرقی	۲۷۶,۱۲۹	۴.۰۶%	۲۸۶,۹۳۳	۴.۰۶%	۲۵۰,۳۵۱	۴.۰۶%	۹۸,۰۸۴	۴.۰۶%	۱۲۳,۸۲۱	۴.۰۶%
آذربایجان‌غربی	۱۳۶,۰۴۳	۲.۰۰%	۱۴۱,۳۶۷	۲.۰۰%	۱۲۳,۳۴۳	۲.۰۰%	۴۸,۳۲۴	۲.۰۰%	۶۱,۰۰۴	۲.۰۰%
اردبیل	۱۳۴,۹۰۶	۱.۹۸%	۱۴۰,۱۸۵	۱.۹۸%	۱۲۲,۳۱۳	۱.۹۸%	۴۷,۹۲۰	۱.۹۸%	۶۰,۴۹۵	۱.۹۸%
اصفهان	۴۹۸,۵۳۵	۷.۳۳%	۵۱۸,۰۴۳	۷.۳۳%	۴۵۱,۹۹۵	۷.۳۳%	۱۷۷,۰۸۵	۷.۳۳%	۲۲۳,۵۵۲	۷.۳۳%
البرز	۱۴۹,۰۰۳	۲.۱۹%	۱۵۴,۸۳۳	۲.۱۹%	۱۳۵,۰۹۳	۲.۱۹%	۵۲,۹۲۷	۲.۱۹%	۶۶,۸۱۶	۲.۱۹%
ایلام	۵۳,۳۴۱	۰.۷۸%	۵۵,۴۲۸	۰.۷۸%	۴۸,۳۶۲	۰.۷۸%	۱۸,۹۴۷	۰.۷۸%	۲۳,۹۱۹	۰.۷۸%
بوشهر	۳۴۲,۶۹۵	۵.۰۴%	۳۵۶,۱۰۴	۵.۰۴%	۳۱۰,۷۰۴	۵.۰۴%	۱۲۱,۷۲۹	۵.۰۴%	۱۵۳,۶۷۱	۵.۰۴%
تهران	۱,۱۷۰,۵۹۶	۱۷.۲۱%	۱,۲۱۶,۴۰۰	۱۷.۲۱%	۱,۰۶۱,۳۱۷	۱۷.۲۱%	۴۱۵,۸۰۸	۱۷.۲۱%	۵۳۴,۹۱۷	۱۷.۲۱%
چهارمحال‌وبختیاری	۵۰,۸۰۶	۰.۷۵%	۵۲,۷۹۴	۰.۷۵%	۴۶,۰۶۳	۰.۷۵%	۱۸,۰۴۷	۰.۷۵%	۲۲,۷۸۳	۰.۷۵%
خراسان‌جنوبی	۷۹,۸۵۰	۱.۱۷%	۸۲,۹۷۴	۱.۱۷%	۷۲,۳۹۵	۱.۱۷%	۲۸,۳۶۳	۱.۱۷%	۳۵,۸۰۶	۱.۱۷%
خراسان‌رضوی	۵۳۹,۸۷۶	۷.۹۴%	۵۶۱,۰۰۱	۷.۹۴%	۴۸۹,۴۷۷	۷.۹۴%	۱۹۱,۷۷۰	۷.۹۴%	۲۴۲,۰۹۰	۷.۹۴%
خراسان‌شمالی	۴۸,۲۸۲	۰.۷۱%	۵۰,۱۷۱	۰.۷۱%	۴۳,۷۷۵	۰.۷۱%	۱۷,۱۵۰	۰.۷۱%	۲۱,۶۵۱	۰.۷۱%
خوزستان	۳۸۱,۶۷۳	۵.۶۱%	۳۹۶,۶۰۸	۵.۶۱%	۳۴۶,۰۴۴	۵.۶۱%	۱۳۵,۵۷۴	۵.۶۱%	۱۷۱,۱۴۹	۵.۶۱%
زنجان	۸۴,۱۷۳	۱.۲۴%	۸۷,۴۶۷	۱.۲۴%	۷۶,۳۱۶	۱.۲۴%	۲۹,۸۹۹	۱.۲۴%	۳۷,۷۴۵	۱.۲۴%
سمنان	۹۶,۶۳۷	۱.۴۲%	۱۰۰,۴۱۹	۱.۴۲%	۸۷,۶۱۶	۱.۴۲%	۳۴,۳۲۷	۱.۴۲%	۴۳,۳۳۴	۱.۴۲%
سیستان‌وبلوچستان	۶۸,۴۲۳	۱.۰۱%	۷۱,۱۰۱	۱.۰۱%	۶۲,۰۳۶	۱.۰۱%	۲۴,۳۰۵	۱.۰۱%	۳۰,۶۸۲	۱.۰۱%
فارس	۵۳۵,۹۹۱	۷.۸۸%	۵۵۶,۹۶۴	۷.۸۸%	۴۸۵,۹۵۵	۷.۸۸%	۱۹۰,۳۹۰	۷.۸۸%	۲۴۰,۳۴۸	۷.۸۸%
قزوین	۱۰۷,۰۴۷	۱.۵۷%	۱۱۱,۲۳۶	۱.۵۷%	۹۷,۰۵۴	۱.۵۷%	۳۸,۰۲۴	۱.۵۷%	۴۸,۰۰۲	۱.۵۷%
قم	۲۲۴,۶۱۸	۳.۳۰%	۲۳۳,۳۰۷	۳.۳۰%	۲۰۳,۶۴۹	۳.۳۰%	۷۹,۷۸۷	۳.۳۰%	۱۰۰,۷۲۳	۳.۳۰%

نسبت‌های ثابت

نمونه‌ای از الگوی آمارسازی ستاد راهیان نور وابسته به ستاد کل نیروهای مسلح

تصویر بالا نمونه‌ای از دستکاری عددسازی نهادهای رسمی ایران را نشان می‌دهد. این آمار را مرکز آمار ایران به نقل از ستاد کل نیروهای مسلح به عنوان آمار شرکت‌کنندگان در اردوهای بازدید از مناطق جنگی (راهیان نور) منتشر کرده است. در نگاه اول همه چیز عادی به نظر می‌رسد، اما دقت در جزئیات اعداد و ارقام نشان می‌دهد، آنچه به عنوان داده ارائه شده، در واقع عددسازی با نسبت‌های ثابت ریاضی است.

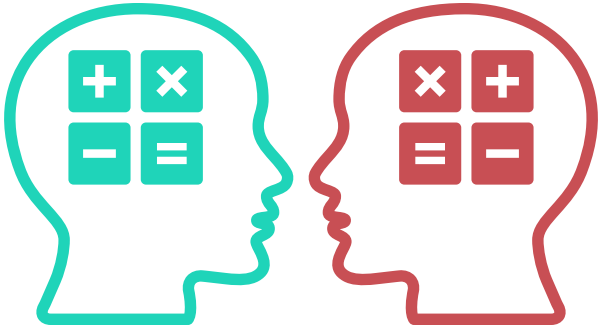
همیشه اما، ایراد از منبع نیست. ممکن است خود ما به **منابع غیرمعتبر** مراجعه کنیم.
یکی از شایع‌ترین انواع این خط ارجاع به گزارش‌های رسانه‌ای به جای منبع اصلی است. بسیاری از خطاهای رایج در جریان استفاده از اطلاعات رسانه‌ها رخ می‌دهد. ممکن است داده‌ها، به‌خصوص اعداد و ارقام، در حین انتقال از منبع اصلی، به عمد یا به سهو تغییر و دستکاری شوند.

۲- اقلام آماری نادرست

ممکن است ما به منابع درست مراجعه کنیم و داده‌های درست در معرض ما باشند، اما ما دو قلم آماری را با هم اشتباه بگیریم.
مثلا اشتباه گرفتن **جمعیت بیکار** به جای **جمعیت غیرفعال** یکی از خطاهای رایج در میان روزنامه‌نگاران است، در حالی که طبق تعریف بیکار کسی است که به رغم آمادگی فعالیت اقتصادی موفق به پیدا کردن کار نشده، اما غیرفعال کسی است که به هر دلیلی تمایل یا آمادگی فعالیت اقتصادی ندارد.

اشتباه‌های محاسباتی

اشتباه‌های محاسباتی یکی از خطراتی که همواره افرادی را که با داده کار می‌کنند تهدید می‌کند. این اشتباه‌ها انواع ساده یا پیچیده‌ای دارند.
اشتباه‌های ساده می‌توانند محاسبات اشتباه ریاضی در حد چهار عمل اصلی باشند، اما نوع دیگری از اشتباه‌های محاسباتی نیز در مراحل تحلیل و پردازش و بازی کردن با داده وجود دارد که می‌توانند یک پژوهش را به بیراهه بکشاند.



به عنوان نمونه می‌توان به استفاده از میانگین هندسی به جای میانگین حسابی اشاره کرد:
برای محاسبه میانگین رشد در یک دوره زمانی باید از میانگین هندسی استفاده کرد، حال آنکه بسیاری از روزنامه‌نگاران و حتی پژوهشگران برای محاسبه متوسط نرخ رشد، از میانگین ساده استفاده می‌کنند.

یک نمونه متداول دیگر از این اشتباه‌ها استفاده بی‌جا از میانه و میانگین است. وقتی با داده‌های پراکنده مواجهیم که فاصله زیادی میان تک داده‌ها وجود دارد، استفاده از میانگین ممکن است گمراه‌کننده باشد. در این شرایط «میانه» تصویر واقع‌گرایانه‌تری از واقعیت ترسیم می‌کند.

خطای پردازش و ارائه

هدف از ارائه متنی و تصویری داده‌ها مانند نمودار و اینفوگرافیک، ارائه تصویری قابل فهم و ملموس از اعداد و ارقام و داده‌ها است برای این کار ما از ابزارهای مفهومی مانند مقایسه استفاده می‌کنیم. مثلا وقتی می‌گوییم نرخ تورم در ایران ۴۵ درصد است، برای درک کردن بزرگی آن باید آن را با نرخ تورم در کشورهای دیگر مقایسه کنیم.

حتی با وجود استفاده از داده‌ای درست و معتبر و محاسبات صحیح، باز هم خطای داده ما را تهدید می‌کند.

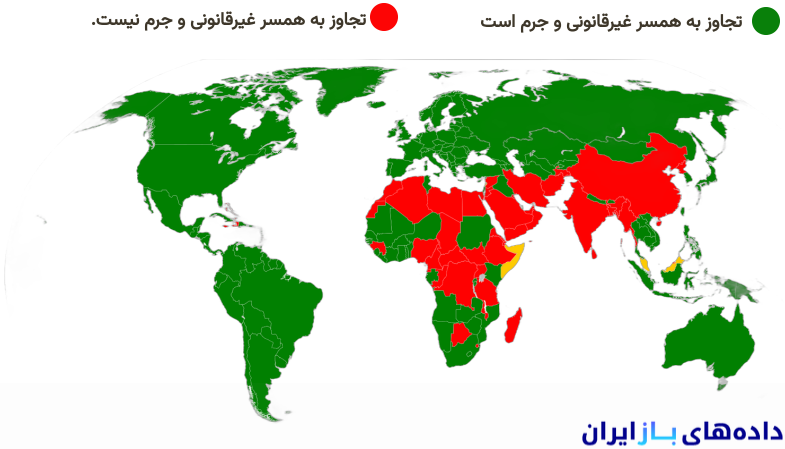
نواع متنوعی از خطای داده هنگام پردازش و ارائه رخ می‌دهد که در اینجا چهار نمونه رایج و متداول آن را بررسی می‌کنیم.

- خطای مقایسه
- خطای مقیاس
- خطای تعدیل
- خطای گزینش

۱- خطای مقایسه

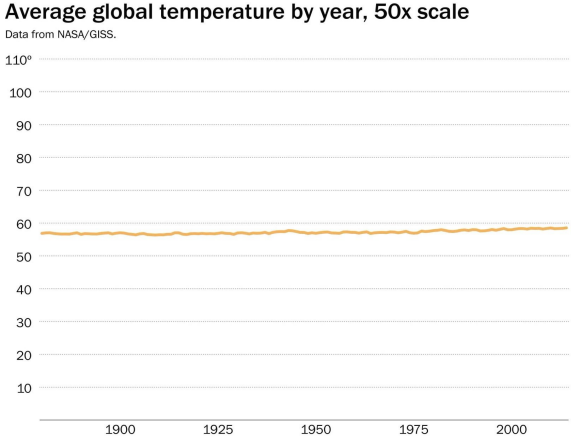
مقایسه وقتی درست است که دو پدیده‌ای که با هم مقایسه می‌شوند باید از نظر منطقی قابل مقایسه باشند. مثلا مقایسه حجم تولید سیب با حجم تولید طلا مقاسه درستی نیست، سیب را باید با گلابی و طلا را مثلا با نقره مقایسه کنیم.
یک نمونه مشهور از خطای مقایسه، گزارش‌هایی است که با استناد به داده‌های معتبر وضعیت خشونت جنسی در کشورها را مقایسه می‌کنند.

به عنوان نمونه اینفوگرافیک بالا در سال ۱۳۹۸ در خبرگزاری مهر تهیه شده است. اما این نمودار بدون اشاره به اینکه مبنای تعریف و گزارش «تجاوز جنسی» در کشورها متفاوت است، مخاطب را دچار اشتباه و بدفهمی می‌کند.
تمام کشورهای بالا کشورهایی هستند که در آنها رابطه جنسی بدون رضایت همسر مصداق تجاوز است و جرم‌انگاری شده، در حالی که در بسیاری از کشورها چنین آماری ذکر نمی‌شود. با این اوصاف نمی‌توان نرخ تجاوز را در سوئد با کشوری مانند ایران مقایسه کرد که چیزی به نام «وادار کردن همسر به رابطه جنسی» را نه‌تنها مصداق تجاوز نمی‌داند، بلکه قانون زنان را مجبور به تمکین از شوهر کرده است.



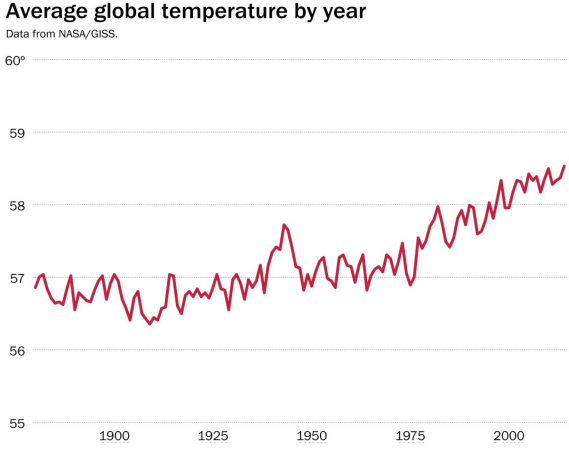
۲- خطای مقیاس

یکی از ابزارهای روزنامه‌نگاری داده‌محور، استفاده از مقیاس برای درک ابعاد پدیده‌ها است. استفاده از مقیاس‌های غیرمنطقی، نه‌تنها به ما برای درک بهتر کمک نمی‌کند، بلکه ما را دچار خطا می‌کند. داده را باید با مقیاس‌های ملموس و منطقی ارزیابی کرد.
یکی از مشهورترین نمونه‌های خطای مقیاس، نمودار مشهور تغییرات درجه حرارت است. هر دو تصویر زیر تغییرات درجه حرارت زمین را بر اساس داده‌های ناسا نشان می‌دهند. اما نمودار اول به بیننده القا می‌کند که بر خلاف گفته‌های دانشمندان درباره گرم شدن زمین، درجه هوا در ۱۳۰ سال گذشته تغییر چندانی نداشته. در حالی که در موضوع گرم شدن زمین دامنه تغییر درجه هوا در یک بازه زمانی ۱۰۰ ساله کمتر از دو درجه است،برای همین باید از مقیاس درست استفاده کرد. چنانکه در نمودار دوم سیر گرم شدن درجه حرارت زمین به طور ملموس و قابل فهم مشخص است.



گمراه‌کننده

مقیاس نامناسب برای نشان دادن تغییرات



درست

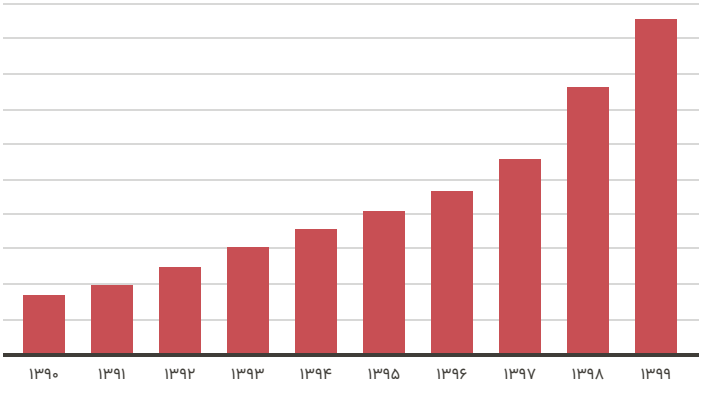
مقیاس مناسب برای نشان دادن تغییرات

۳- خطای تعدیل

وقتی می‌خواهیم پدیده‌ها را در طول زمان با هم مقایسه کنیم باید مراقب به تغییراتی باشیم که در دوره‌های زمانی داده‌ها را تحت تاثیر قرار می‌دهد.
این مساله به طور مشخص در مقایسه قیمت اجناس به خصوص در کشورهایی که میزان تغییر قیمت‌ها و تورم در آنها بالا است بسیار حیاتی است.

گمراه‌کننده

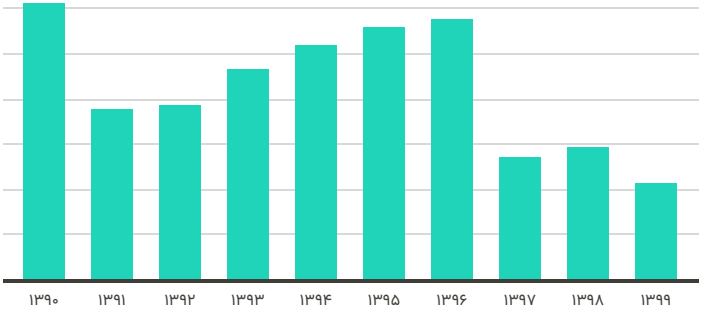
ارزش اسمی حداقل دستمزد کارگران بدون تعدیل



مثلا مقایسه دستمزد کارگران در سال‌های دهه ۱۳۹۰ نشان می‌دهد که ارزش اسمی حداقل دستمزد کارگران ۶ برابر شده است (نمودار بالا) اما اگر ارزش واقعی دستمزد را با احتساب یک قیمت ثابت مثلا ارزش دلار در بازار آزاد مقایسه کنیم (نمودار پایین) می‌بینیم نه‌تنها ارزش حقوق کارگران بیشتر نشده، بلکه به کمتر از یک سوم ۱۰ سال قبل سقوط کرده است.

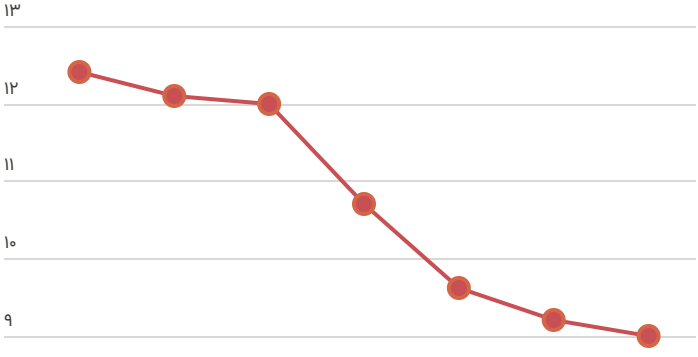
درست

ارزش حداقل دستمزد کارگران تعدیل با قیمت دلار



۴- خطای گزینش

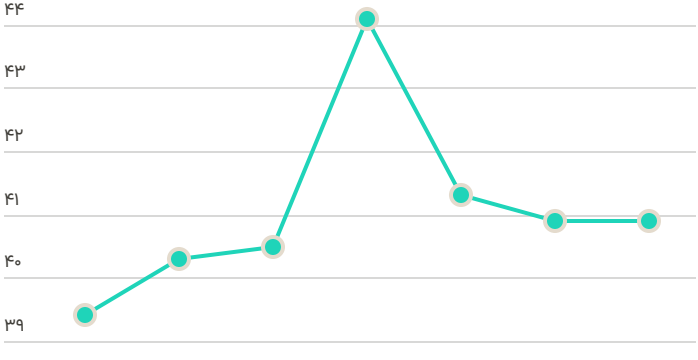
در مقابل ما داده‌های زیادی وجود دارند که استفاده گزینشی از هر کدام از آنها می‌تواند تصویری متفاوت از واقعیت ترسیم کند. حتی گاهی استفاده گزینشی از شاخص‌های متداولی مانند نرخ بیکاری می‌تواند گمراه‌کننده باشد. به عنوان مثال مقایسه نرخ بیکاری در ۵ سال اخیر نشان از کاهش بیکاری دارد، اما آیا این به معنای بهتر شدن بازار کار است؟



گمراه‌کننده

نرخ بیکاری

خیر. در همین مدت باز کار در ایران نه‌تنها بهتر نشده بلکه کوچک هم شده است. در واقع کاهش نرخ بیکاری هم نه نتیجه بهبود شرایط اقتصادی، بلکه نتیجه بحرانی است که در نتیجه خروج افراد از باز کار کاهش جمعیت فعال اقتصادی به وجود آمده است.
در واقعیت اگر ما همزمان با مقایسه نرخ بیکاری، نرخ مشارکت اقتصادی را هم در نظر بگیریم، می‌توانیم تصور کامل‌تر و دقیق‌تری از واقعیت داشته باشیم.



درست

نرخ مشارکت اقتصادی