* آزمون فرضیه در مقابل تجزیه و تحلیل داده های اکتشافی

هنگامی که به یک مشکل داده کاوی بر میخوریم ، یک تحلیلگر داده کاوی ممکن است از قبل فرضیه های پیشینی داشته باشد که بخواهد در رابطه با روابط بین متغیرها آزمایش کند. به عنوان مثال، فرض کنید که مدیران شبکه های تلفن همراه علاقه مند هستند که آیا افزایش اخیر در ساختار هزینه منجر به کاهش استفاده از تلفن شده است یا خیر. در این مورد، تحلیلگر این فرضیه را که استفاده از تلفن کاهش یافته است آزمایش می کند و بنابراین از روش های آزمون فرضیه استفاده می کند.

تعداد بی شماری از روش های آزمون فرضیه های آماری در تحلیل آماری سنتی موجود است. ما بسیاری از این موارد را در فصول 4 و 5 پوشش می دهیم. با این حال، تحلیلگران همیشه مفاهیم پیشینی از روابط مورد انتظار بین متغیرها ندارند. به ویژه هنگامی که با پایگاه داده های ناشناخته و بزرگ مواجه می شوند، تحلیلگران اغلب ترجیح می دهند از تجزیه و تحلیل داده های اکتشافی (EDA) یا تجزیه و تحلیل داده های گرافیکی استفاده کنند. EDA به تحلیلگر اجازه می دهد

* کاوش در مجموعه داده ها
* بررسی روابط متقابل بین صفات؛
* زیر مجموعه های مورد نظر در مشاهدات را شناسایی کنید. و
* یک ایده اولیه از ارتباط های احتمالی بین پیش بینی کننده ها و همچنین بین پیش بینی کننده ها و متغیر هدف ایجاد کنید.
* آشنایی با مجموعه داده ها

گراف ها، نمودارها و جداول اغلب روابط مهمی را آشکار می کنند که می تواند زمینه های مهمی را برای بررسی بیشتر نشان دهد. در این فصل، ما از روش‌های اکتشافی برای بررسی مجموعه داده‌های Churn استفاده میکنیم. Churn که فرسایش نیز نامیده می شود، اصطلاحی است که برای نشان دادن ترک کردن خدمات یک کمپانی از طرف مشتری و روی آوردن به شرکت دیگر استفاده می شود. مجموعه داده شامل 20 پیش‌بینی‌کننده اطلاعات مربوط به 3333 مشتری، همراه با متغیر هدف، ریزش است، که نشان می‌دهد آیا آن مشتری از شرکت خارج شده یا نه.