به نام خدا

دانشگاه اصفهان

گزارش اول داده کاوی

استاد مربوطه : دکتر غلامی

پرستو غلامی

۹۹۳۶۱۳۰۴۸

دی ۱۴۰۲

در ابتدا دو کتابخانه pandas برای کار با داده ها و matplotlib برای رسم نمودار ایمپورت میکنیم.
پس از خواندن فایلcsv ، چاپ شده آنرا میتوانید مشاهده کنید.
تابعstatistic
خروجی این تابع بترتیب تعداد کلمات به کار رفته، تعداد داده های گم شده، بیشترین کلمه تکرار شده و تکرار ان، کمترین کلمه تکرار شده و تکرار ان میباشد که در دیتا فریم statics ذخیره میشود.
همچنین کلمات به کار رفته در هر ستون مشخص شده با از کامنت دراوردن چاپ value تمام کلمات هر سطر با تکرار انها چاپ میشود.
سپس تغییرات تکمیلی را برای اینکه دیتافریم static تایپ را نمایش دهد اعمال شده است.
در قسمت بعد برای هر ستون دلخواه میتوان جدولی برای نمایش کلمات و تکرار انها رسم کرد.
با تابع describe میتوان اطلاعاتی مثل تعداد، میانگین، انحراف معیار،کمترین مقدار، چارک اول و دوم و سوم، بیشترین مقدار را از ستون هایی که مقدار عددی دارند دریافت کرد.
در ادامه مدیریت داده ناسازگار را داریم که همانطور که میبینید داده ۹۹ را به nan تغییر داده ایم تا در ادامه با یکی از روش های ذکر شده داده های گم شده را پر کنیم.
پر کردن مقادیر داده های گم شده از ۴ طریق میانگین، میانه، مد و حذف رکورد ممکن است که هر ۴ روش آمده است.
همچنین کد تقسیم دیتا فریم به دیتا فریم کوچکتر و ساخت زیر مجموعه از آن را میتوانید مشاهده کنید.