گزارش تمرین اول دوره‌ی یادگیری ماشین

سجاد عابد

در سلول اول کتابخانه‌های مورد نیاز فراخوانی می‌شوند.

در سلول دوم فایل csv توسط کتابخانه‌ی با فرمت DataFrame در داخل متغیر data ریخته می‌شود.

در سلول سوم shape و ۵ سطر اول جدول نمایش داده می‌شوند. در اینجا می‌فهمیم که جدول ما ۱۳ ستون و ۱۰۸۴۱ رکورد (ردیف) دارد و نام ستون‌ها مشخص می‌شود.

در سلول چهارم infoی جدول را دریافت می‌کنیم. مشاهده می‌شود که فقط یکی از ستون‌ها از نوع عددی است. همچنین از ستون Non-Null Count می‌توان متوجه شد که کدام ستون‌ها داده‌ی خالی دارند. با توجه به اینکه ستون داده‌های Price به راحتی قابل تبدیل به مقادیر عددی هستند و در این صورت می‌توان تحلیل بهتری روی آن‌ها انجام داد، در سلول شش آن‌ها را به مقادیر عددی تبدیل می‌کنیم. در سلول پنج مقادیر خاص Price را بررسی کردیم که دیدیم تنها یک ردیف (۱۰۴۷۲) از فرمت دیگر داده‌ها پیروی نکرده است.

در سلول ششم ردیف‌هایی که داده‌ی 0 یا Everyone دارند، مقدار 0 می‌گیرند و باقی ردیف‌ها پس از کنار گذاشتن علامت $ باقی عدد را به صورت float دریافت می‌کنند.

همین کار را در مورد ستون Reviews انجام دادیم.

در سلول ۹ دوباره ردیف‌های بالایی جدول را مشاهده می‌کنیم که در این جدول داده‌های ستون Price به روزرسانی شده اند.

در سلول 10 نیز خلاصه‌ی داده‌های جدول گزارش می‌شود. با توجه به اینکه بیشتر داده‌ها عددی نیستند، بسیاری از داده‌های آماری در مورد آن‌ها قابل ارائه نیستند.

در سلول ۱۱ کراستب بین نوع برنامه Content Rating آن‌ها داده شده که داده‌ها بر اساس ستون‌ها نرمالایز شده اند.

در سلول ۱۲ نمودار قیمت داده‌ها نمایش داده شده است که بر اساس آن می‌توان فهمید که مقدار خیلی کمی از برنامه‌ها هزینه‌ی حدود ۲۰ دلار و تعداد کمی نیز هزینه‌ی نزدیک ۴۰۰ دلار دارند و از حدود ۲۰ تا حدود ۴۰۰ دلار تقریبا هیچ برنامه‌ای وجود ندارد.

در ادامه در سلول ۱۴ می‌توان پراکندگی Category داده‌ها بر اساس Content Rating آن‌ها مشاهده کرد. همچنین از طریق رنگ دایره‌ها می‌توان فهمید که میزان نصب آن‌ها به چه میزان بوده است.

در سلول ۱۵ نیز نمودار Reviews بر اساس Sizeرسم شده است که بر اساس شکل و رنگ نقطه‌ها می‌توان فهمید که مربوط به کدام Type و Content Rating هستند.

در سلول آخر نیز مشخص است که برای هر content Rating نمودار قیمت و میانگین امتیازات چگونه بوده است. از این نمودار در اولین نگاه می‌توان این نتیجه را گرفت که بیشتر برنامه‌های پولی، امتیاز بالای ۳ کسب کرده اند و تعداد برنامه‌های پولی که نمرات پایینی دریافت کرده اند بسیار ناچیز است.

برای بررسی بیشتر در سلول ۱۷ نمودار فقط برای برنامه‌هایی که باید برای نصب آن‌ها هزینه‌ای پرداخت کرد، نمایش داده می‌شوند.