در بخش اول کتابخانه های مورد نیاز اضافه شدند.

در بخش دوم بر اساس داده بارگزاری شده یک دیتافریم با نام google\_df ایجاد شد.

در بخش سوم که به کشف ویژگی های دیتافریم اختصاص یافته بود، موارد متعددی چک شد:  
ابتدا ساختار کلی دیتافریم و سطر و ستون ها بررسی شد. ابعاد دیتافریم و نام ستون ها استخراج شد. تعداد هر ستون چاپ شد که می تواند برای یافتن missing value راهنمای مناسبی باشد. Type هر کدام از ستون ها هم بررسی شد. خلاصه ای از موارد هم با دستور info نمایش داده شد. در قسمت دیگر هم تعداد مقادیر unique ستون های مختلف بررسی شد. در برخی از دستورات با توجه به اینکه jupyter خروجی آخرین دستور را نمایش می دهد دیگر از دستور print استفاده نشده است.

در دیتاست استفاده شده برای این تمرین مقادیر متفاوتی از missing value وجود دارد. مهم ترین آن در بخش rating است که براساس خروجی دستورات مشخص است که تعداد داده کمتر از سایر ستون ها دارد. در خصوص رفتار با missing value در برخی موارد ممکن است که میانگین گیری کمک کننده باشد. در این جا با توجه به استقلال اپلیکیشن ها و عدم تاثیر مستقیم سایر موارد بر Rating به نظر استفاده از میانگین گیری کمک کننده نخواهد بود. البته اگر اپلیکیشن ها را به لحاظ میزان نصب دسته بندی کنیم شاید بتوان بر اساس آن rating ای برای برخی اپلیکیشن ها به خصوص برای اپلیکیشن های با تعداد دفعات بالا در نظر گرفت. اما در خصوص اپلیکیشن های با دانلود پایین با توجه به محدوده گسترده تغییرات این روش پاسخگو احتمالا دقیق نخواهد بود. از طرف دیگر در بخش Size هم مقدار varies with device دارد که در تحلیل کارایی خاصی نخواهد داشت. به همین دلیل یک دیتافریم دیگر هم با حذف سطور دارای این مقادیر ساخته شد که در برخی تحلیل ها مورد استفاده قرار بگیرد. هم چنین برای تحلیل size مقادیر چون به مگابایت و کیلو بایت نوشته شده بود همه به صورت کیلوبایت تبدیل شد.(ضرب ۱۰۲۴ در فایل ها )

در بخش visualizing از دو دیتافریم google\_df و google\_df\_size استفاده شده که در دومی مقادیر سایز بر اساس کیلوبایت اصلاح شده است.

براساس پارامترهای مختلفی که وجود داشت نمودارهای مختلفی رسم شده است. برخی از نتایجی که از این نمودارها به دست می آید:

همان طور که پیش بینی می شد تعداد نصب نرم افزارهای رایگان از نرم افزاری پولی بیشتر بود اما نمودار review آن ها تقریبا به هم شباهت دارد. این طور می توان استدلال کرد که زمانی که یک نرم افزار متوسط و خوب است افراد بیشتری آن را نصب می کنند و به همین منوال نظرات بیشتری در مورد آن ها داده می شود از طرف دیگر نرم افزارهایی که نمرات کمتری دارند تعداد افراد کمتری آن را دانلود و به تبع آن نظرات کمتری دارند. نکته دیگر این است که نرم افزارهای با تعداد نصب بسیار بالا دارای نمرات بالاتر از میانگین است که می تواند در اثر تاثیر نمره بالا در نصب سایر افراد باشد. یعنی افراد بر اساس نمرات یک نرم افزار تصمیم می گیرند که آن را نصب کنند یا نه. از طرفی سایز نرم افزار هم در تعداد نصب تاثیر گذار بوده است. نرم افزارهای کم حجم تعداد بیشتری دانلود شده اند( البته ممکن است این به خاطر نمونه داده باشد که شامل نرم افزارهای حجیم نیست). در یک بخش هم برای اینکه نمودار مناسب تری بر اساس تعداد نصب باشد نرم افزارهایی که بسیار نصب شده اند را از داده حذف کردم تا اطلاعات بهتری نسبت به نرم افزارهایی که کمتر نصب شده اند به دست بیاید(با عنوان google\_df\_less) . بخشی هم به بررسی تعداد نصب و rating بر اساس دسته بندی نرم افزار پرداختم که نشان می داد که نرم افزارهایی در دسته game و lifestyle و shopping و .. نرم افزارهایی هستند که بازخورد مناسب تری داشته اند و نمره بالاتری هم دریافت کرده اند اما به لحاظ نصب شاید نتوان به راحتی نظر داد(البته با جداسازی می توان نمودارهایی بهتری هم رسم کرد.) به طور خلاصه سعی شد بر اساس پارامترهای ratingو installs و size و review و استفاده از type و category درک ابتدایی از داده ها کسب کرد.