**مقدمه**

EDA [[1]](#footnote-1)رویکردی برای تحلیل داده‌ها است. فلسفه پشت این رویکرد بررسی داده ها قبل از ساخت یک مدل بود. اگر تحلیل داده‌ها درست صورت بگیرد، در بهبود انجام روند مدل‌سازی نیز می‌تواند موثر باشد. EDA برای دریافت اطلاعات و فهم بهتر داده ها استفاده می شود. دانشمندان و تحلیلگران داده سعی می کنند با استفاده از نمودارهای آماری و سایر تکنیک های تجسم، الگوها، روابط و ناهنجاری های مختلف را در داده ها بیابند.

هدف اصلی EDA تشخیص هر گونه خطا، نقاط پرت و همچنین درک الگوهای مختلف در داده ها است. این به تحلیلگران اجازه می دهد تا قبل از هر گونه فرضی، داده ها را بهتر درک کنند. این روش نه تنها بینشی در مورد داده ها می دهد، بلکه مجموعه داده ای را نیز برای توسعه و مدل‌سازی مناسب به ما می دهد. نتایج EDA به کسب و کارها کمک می کند تا مشتریان خود را بشناسند، تجارت خود را گسترش دهند و بر اساس آن تصمیم بگیرند. این مسئله برای هر کسب و کاری مفید است، زیرا اغلب با بینش های غیرقابل پیش بینی مواجه می شوند. در نهایت فرقی نمی‌کند که هدف کسب و کار‌ها و شرکت ها چه باشد ، اولین قدم همیشه با یک EDA شروع می شود. این یکی از اجزای مهم فرآیند تحقیقات بازاریابی است که به داده ها اجازه می دهد تا به نفع کسب و کار شما سازماندهی، بررسی و تفسیر شوند.

روش ها و هدف های مختلفی در ارتباط با کار و تحلیل داده ها وجود دارد ؛ از بررسی داده ها برای پیدا کردن خطاها و داده های ناقص تا نگاهی به روابط بین متغیر‌ها و اهیمت روند حرکت داده ها در طول زمان .

مصور‌سازی داده ها جزء مهمی از تجزیه و تحلیل داده های اکتشافی ( (EDAاست زیرا به تحلیلگراجازه می دهد تا به داده های خود نگاه کند و متغیرها و روابط بین آنها را بشناسد.

در انجام این پروژه از نمودارهایی از جمله هیستوگرام ، باکس پلات ، بار چارت و نمودار دایره‌ای استفاده شده که در ادامه به صورت مختصر به معرفی آن‌ها می‌پردازیم.

* **هیستوگرام [[2]](#footnote-2) :** نمودارهای هیستوگرام برای نشان دادن توزیع هر متغیر پیوسته و میزان فراوانی آن استفاده می شود.
* **نمودار جعبه ای [[3]](#footnote-3) :** روشی استاندارد برای نمایش توزیع داده ها بر اساس یک خلاصه پنج عددی ("کمترین"، چارک اول ( Q1)، میانه، چارک سوم ( ( Q3 و "بیشترین") است. این می تواند به شما در مورد نقاط پرت و ارزش آنها بگوید.
* **نمودار دایره ای [[4]](#footnote-4) :** داده ها را به صورت درصدی از کل نشان می دهد.
* **نقشه حرارتی [[5]](#footnote-5) :** نقشه های حرارتی برای مشاهده همبستگی بین متغیرهای ویژگی استفاده می شود. این امر به ویژه زمانی مهم است که ما در تلاش برای به دست آوردن اهمیت ویژگی در تحلیل رگرسیون هستیم. اگرچه ویژگی‌های مرتبط بر عملکرد مدل آماری تأثیر نمی‌گذارد، اما می‌تواند تحلیل پس از مدل‌سازی را مختل کند.
* **نمودار Q-Q :** در آمار این نمودار جز نمودار احتمال است، یک روش گرافیکی برای مقایسه دو توزیع احتمال با رسم چندک های آنها در برابر یکدیگر.

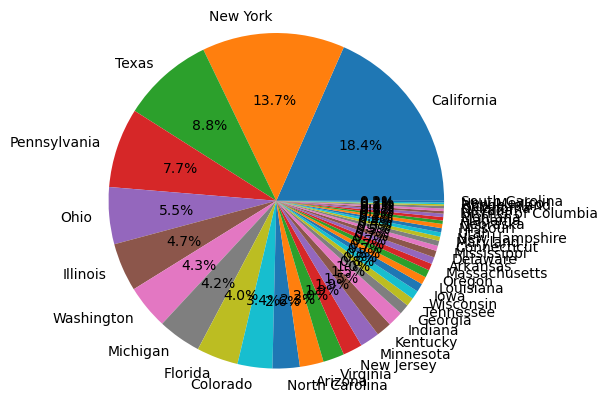
داده های در دسترس مربوط به یک فروشگاه خرده فروشی است که با لوازم مصرفی، شرکتی و اداری خانگی سروکار دارد. فروش به صورت آنلاین انجام می شود و حمل و نقل به تمام ایالت های ایالات متحده ارائه می شود. هدف از این پروژه تجزیه و تحلیل فروش ، بررسی سود و زیان در ارتباط با محصولات مختلف و به طور کلی ارزیابی بخش های مختلف فروشگاه و به تصویر کشیدن اطلاعاتی است که به خودی خود ممکن است خیلی قابل لمس و تشخیص نباشند. بنابراین تجزیه و تحلیل دیدگاهی در مورد روند کسب و کار ارائه می دهد. با تقاضاهای رو به رشد و رقابت‌های بی‌نظیر در بازار این برای کسب و کار ضروری و قابل اهیمت است که چیزی برای آنها بهترین است آیا مسیر خود را حفظ کنند یا تغییراتی را برای بهینه سازی فروش ایجاد کند. با توجه به نتایج به‌دست‌آمده از EDA، فروشگاه می‌تواند محصولات غیر سودآور را حذف کند یا در تلاش‌های بازاریابی برای محصولات، بخش‌ها و مناطق جغرافیایی سرمایه‌گذاری کند که منجر به سود آنها می‌شود.

در ابتدا به دنبال همبستگی ها خواهیم بود و برای این کار از نقشه حرارتی ستون های مختلف را با یکدیگر مقایسه می کنیم.



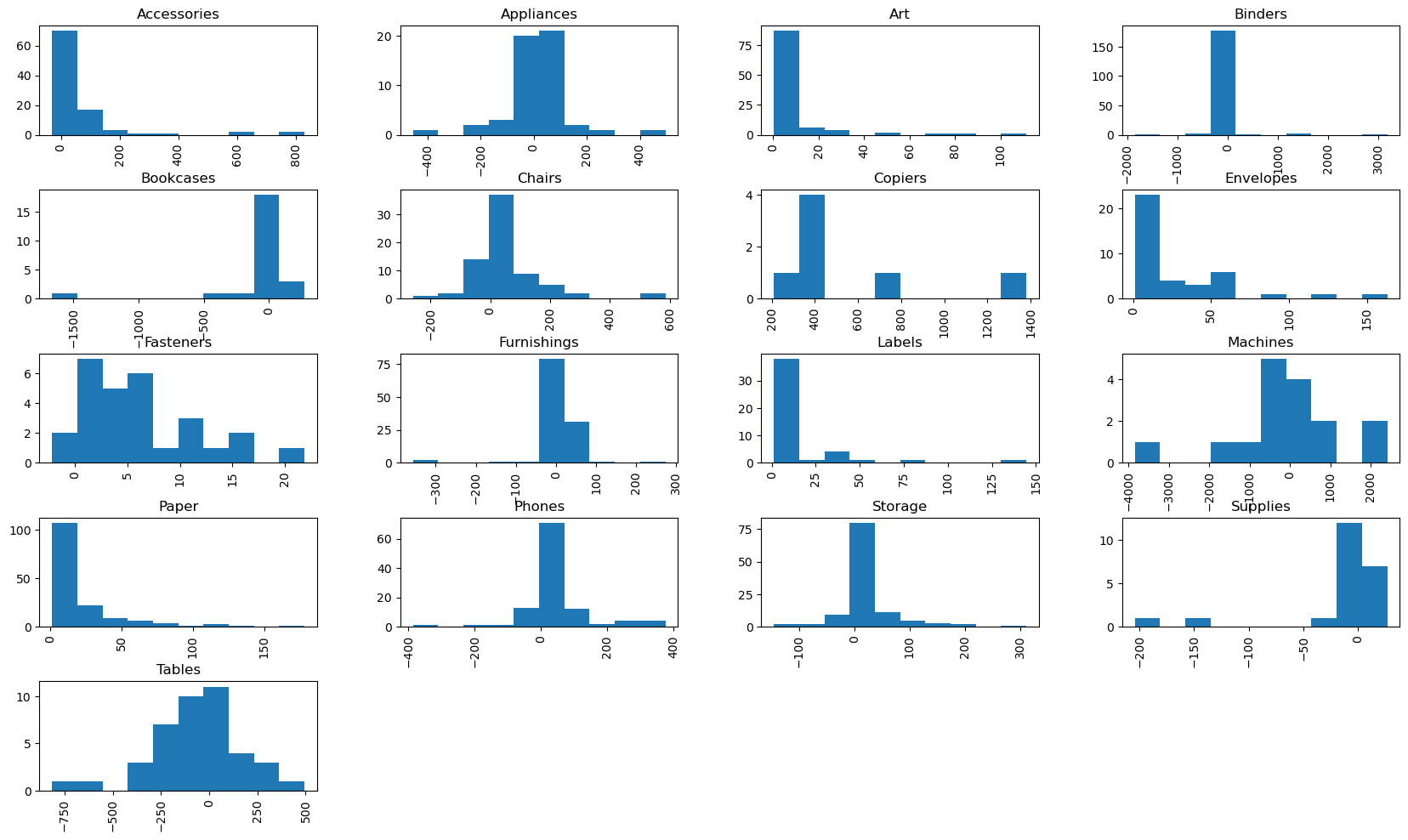
شکل ۱- نقشه حرارتی با استفاده از پارامتر همبستگی

همانطور که در شکل ۱ مشخص است ، بیشترین همبستگی بین ستون های هزینه تولید و فروش برقرار است.



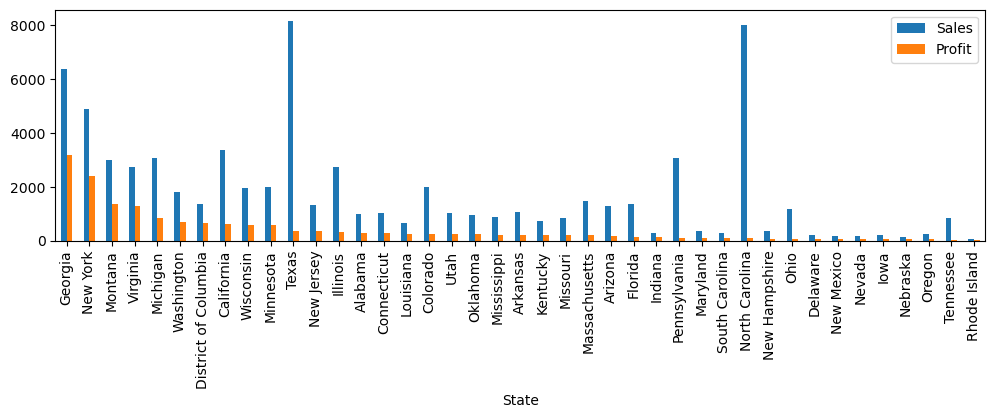
شکل ۲- نمودار دایره ای

شکل ۲ نمایانگر ایالت ها بر اساس بیشترین **تعداد** فروش محصول است.

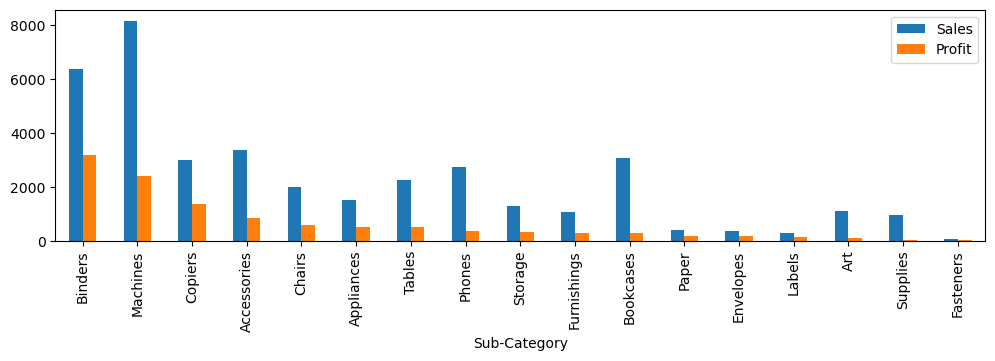


شکل۳- نمودار فراوانی در هر زیرگروه محصول

در نمودار های شکل فراوانی و توزیع سود هرمحصول قابل مشاهده می‌باشد.



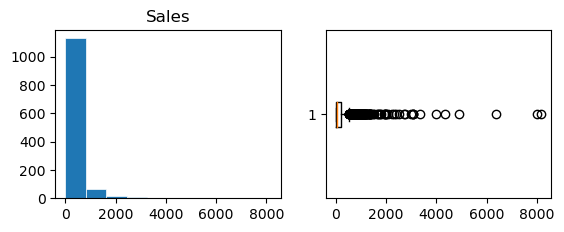
شکل ۴- نمودار میله ای

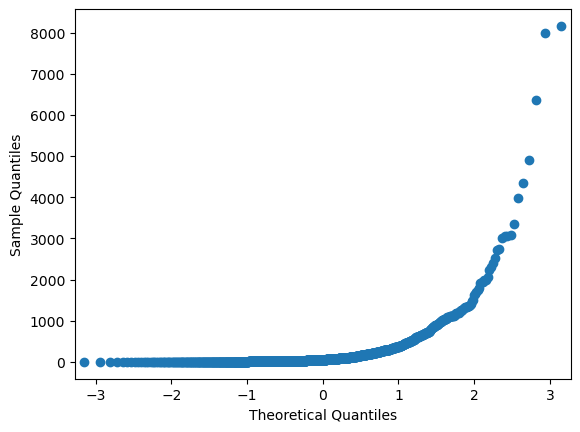


شکل ۵

برای نشان دادن پرسود ترین محصولات و سودآورین ترین ایالت ها از نمودار میله ای در شکل ۴ و ۵ استفاده کرده ایم .

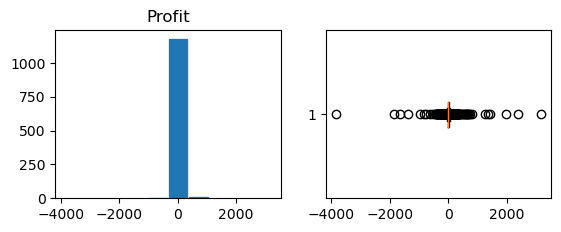
نمودارهای جعبه ای (یا نمودارهای ویولن) و هیستوگرام ها انتخاب های رایجی برای خلاصه سازی بصری یک متغیر کمی هستند. این نمودارها به این دلیل مفید هستند که به طور همزمان اطلاعاتی در مورد مقادیر حداقل و حداکثر، مکان مرکزی و گسترش ارتباط برقرار می کنند. هیستوگرام ها علاوه بر این می توانند الگوهایی را که می توانند بر تجزیه و تحلیل تأثیر بگذارند (به عنوان مثال، کج یا چولگی داشتن) روشن کند.

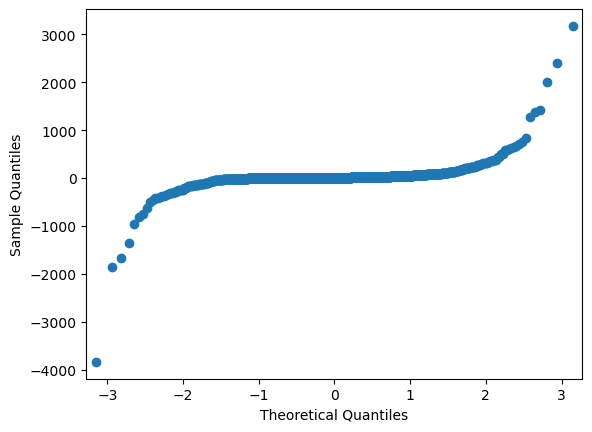




شکل ۶

همانطور که در شکل۶ مشخص است ستون فروش در نمودار هیستوگرام دارای چولگی به سمت راست (مثبت ) است . زمانی که دنباله راست نمودار هیستوگرام توزیع طولانی تر است و اکثر مشاهدات در دنباله چپ متمرکز شده است. چولگی مثبت، چولگی راست یا دنباله راست نامیده می شود. در این حالت میانه کوچکتر از میانگین است.





شکل۷

**فهرست منابع**

***Nimit Vanawat, How To Perform Exploratory Data Analysis. Published On August 12, 2021 and Last Modified On July 26th, 2022 Available from: www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/***

***Jiahao Weng , Exploratory Data Analysis . Published On Sep 24, 2019. Available from: towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide-and-template-for-structured-data-abfbf3ee3bd9***

# ***Samantha Lile, Types of Graphs Perfect for Every Top Industry, Available from:visme.co/blog/types-of-graphs***

1. Exploratory data analysis [↑](#footnote-ref-1)
2. Histogram [↑](#footnote-ref-2)
3. Boxplot [↑](#footnote-ref-3)
4. Pie chart [↑](#footnote-ref-4)
5. Heatmap [↑](#footnote-ref-5)