به نام خدا

گزارش پروژه میان ترم علم داده

تحلیل اکتشافی داده (EDA) بر روی دیتاست Absenteeism time in hours

نیایش خانی ۹۹۵۲۱۲۳۵

استاد درس: دکتر نادری

نیم سال اول ۱۴۰۴-۱۴۰۳

چکیده

این پروژه به بررسی مجموعه داده‌ای از برزیل می‌پردازد که در آن سوابق غیبت در محل کار از ژوئیه ۲۰۰۷ تا ژوئیه ۲۰۱۰ در یک شرکت پیک جمع‌آوری شده است. هدف این پروژه کاهش غیبت کارمندان و بررسی علل اولیه غیبت آنها است.
همچنین در این پروژه به این سوالات پاسخ داده می‌شود:
- کدام بخش از زندگی (مانند کار، خانواده و ...) بر غیبت تأثیر می گذارد؟
- آیا بین دلیل غیبت و غیبت کردن رابطه آشکاری وجود دارد؟

نکته: موارد خط‌کشی شده به همراه باقی مغایرت‌ها با فایل اصلی، توسط اینجانب اضافه شده اند.

مقدمه

هر محیط کاری، فارغ از ساختار و اندازه، با چالش غیبت کارکنان روبرو است. دلایل این پدیده متنوع و گاه پیچیده هستند. غیبت، چه به صورت برنامه‌ریزی شده و چه ناگهانی، می‌تواند تأثیرات قابل توجهی بر بهره‌وری، روحیه سازمانی و حتی عملکرد مالی یک شرکت داشته باشد.
همان‌طور که مجله Forbes اشاره می‌کند، غیبت مکرر می‌تواند به کاهش چشمگیر بهره‌وری منجر شود. دلایل این پدیده گسترده از بیماری‌های ساده تا مشکلات خانوادگی و حتی نارضایتی شغلی متغیر است.

در این پژوهش، با تمرکز بر داده‌های یک شرکت پیک در برزیل از سال ۲۰۰۷ تا ۲۰۱۰، به دنبال یافتن ریشه‌های اصلی غیبت و راهکارهایی برای کاهش آن هستیم. هدف اصلی، درک عمیق‌تر از عوامل موثر بر غیبت و شناسایی ارتباط احتمالی بین دلایل مختلف غیبت و میزان آن است.

سوالات کلیدی که در این پژوهش مورد بررسی قرار می‌گیرند:

- **تعادل کار و زندگی:** کدام جنبه از زندگی کارکنان (کار، خانواده، زندگی شخصی) تأثیر بیشتری بر تصمیم به غیبت دارد؟
- **تأثیر دلایل غیبت:** آیا بین دلایل مختلف غیبت (مانند بیماری، مشکلات خانوادگی، نارضایتی شغلی) و میزان غیبت ارتباط مستقیمی وجود دارد؟
- **عوامل محیطی:** آیا عوامل محیطی مانند نوع شغل، شرایط کاری، و فرهنگ سازمانی بر میزان غیبت تأثیرگذار هستند؟
- **الگوهای غیبت:** آیا الگوهای مشخصی در غیبت کارکنان وجود دارد (مثلاً غیبت‌های متناوب، غیبت‌های طولانی‌مدت)؟

با پاسخ به این سوالات، می‌توانیم به درک جامع‌تری از پدیده غیبت دست پیدا کنیم و در نهایت، راهکارهای عملی برای کاهش آن ارائه دهیم. نتایج این پژوهش می‌تواند برای مدیران منابع انسانی، سیاست‌گذاران و محققان حوزه مدیریت منابع انسانی مفید باشد.

کاوش داده‌ها

این مجموعه داده شامل ۷٤۰ رکورد و ۲۱ ویژگی با موارد زیر است:

- ۸ فیچر دسته‌بندی:

- Reason for absence
- Month of Absence
- Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
- Seasons
- Disciplinary failure(yes=1; no=0)
- Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
- Social drinker (yes=1; no=0)
- Social smoker (yes=1; no=0)

- ۱۲ فیچر عددی:

- ID
- Transportation Expense
- Distance from Residence to Work(kilometers)
- Service time
- Age
- Work Load
- Hit target
- Son (number of children)
- Pet (number of pet)
- Weight
- Height
- Body mass index

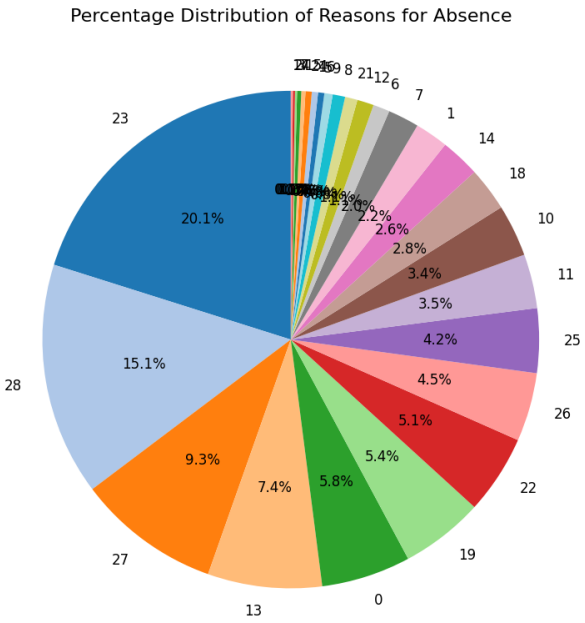
- متغیر تارگت: Absenteeism time in hours

فیچر دلایل غیبت (Reasons for absence) به ۲۱ دسته تقسیم‌بندی شده است:

1. Certain infectious and parasitic diseases	2. Neoplasms
3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	4. Endocrine, nutritional and metabolic diseases
5. Mental and behavioural disorders	6. Diseases of the nervous system
7. Diseases of the eye and adnexa	8. Diseases of the ear and mastoid process
9. Diseases of the circulatory system	10. Diseases of the respiratory system
11. Diseases of the digestive system	12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue	14. Diseases of the genitourinary system
15. Pregnancy, childbirth and the puerperium	16. Certain conditions originating in the perinatal period
17. Congenital malformations, deformations and chromosomal abnormalities	18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
19. Injury, poisoning and certain other consequences of external causes	20. External causes of morbidity and mortality
21. Factors influencing health status and contact with health services.	22. Patient follow-up
23. Medical consultation	24. Blood donation
25. Laboratory examination	26. Unjustified absence
27. Physiotherapy	28. Dental consultation

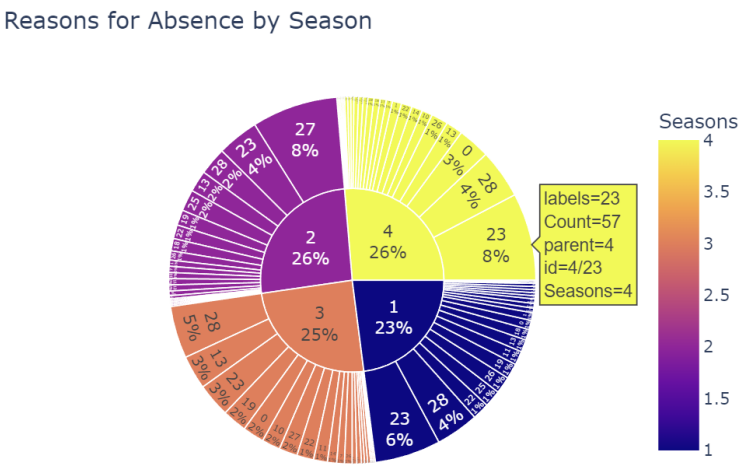
Visualize کردن تعدادی از فیچرها:

برای یافتن اینکه کدام یک از دلایل غیبت بیشتر تکرار شده است، می‌توانیم از pie chart استفاده کنیم.



همانطور که مشاهده می‌شود دلیل شماره ۲۳ یعنی Medical consultation و بعد از آن دلیل ۲۸، Dental consultation، بیشترین تکرار را داشته است.

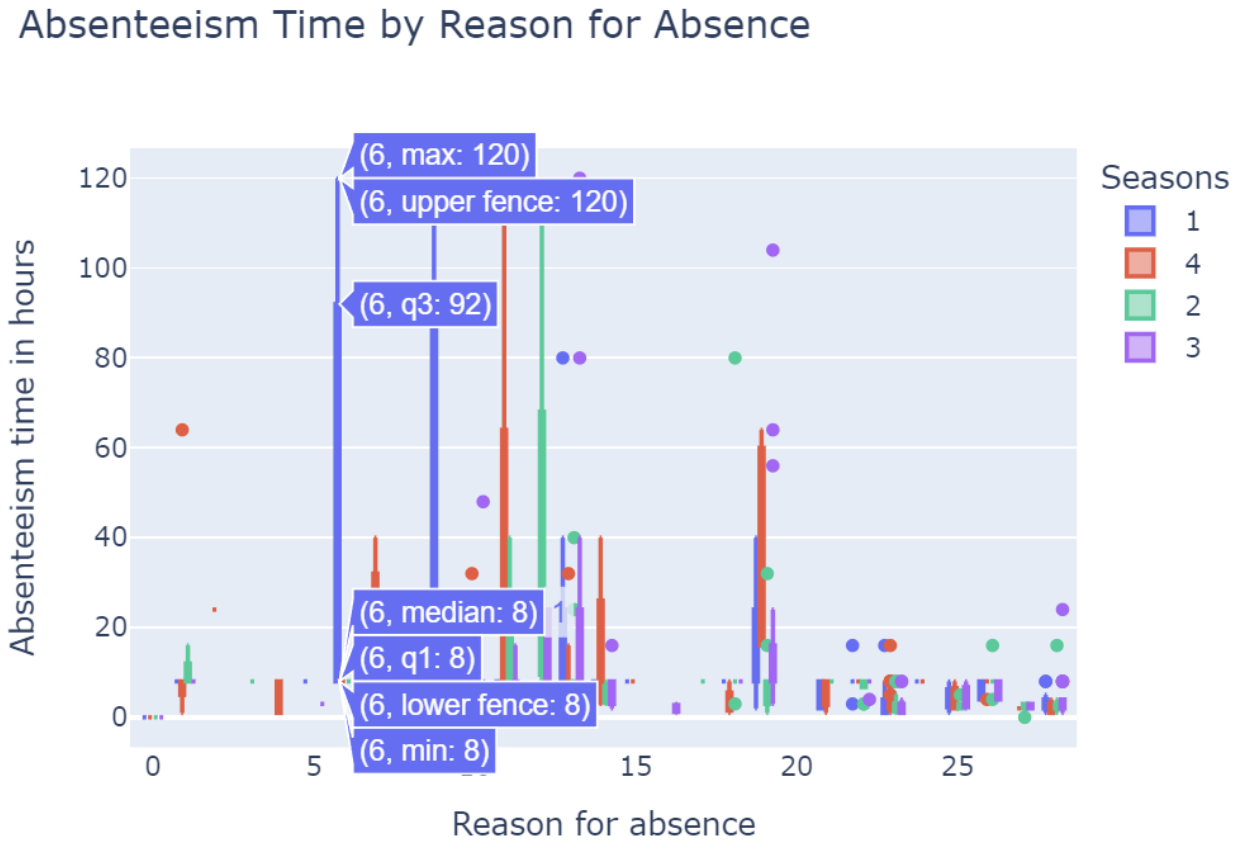
همچنین می‌توان بررسی کرد که در هر فصل کدام یک از دلایل غیبت بیشترین تکرار را داشته است.



تصویر گویا است که در فصل ۴، پر تکرارترین دلیل غیبت دلیل ۲۳ است که ۵۷ بار تکرار شده است.

Visualize کردن توزیع ساعات غیبت بر دلایل غیبت

می‌توانیم به وسیله نمودار زیر دریابیم که بیشترین میزان ساعت غیبت با کدام دلایل رخ داده است.



برای مثال بیشترین ساعات غیبت با دلیل شماره ۶ رخ داده است.

پیش پردازش

به منظور ساخت یک مدل خوب، درک کمی از مجموعه داده برای نتایج بهتر مورد نیاز است.
ابتدا یک تحلیل کلی با آمار توصیفی انجام می‌دهیم تا شکل یا tendency هر ستون را به تصویر بکشیم.
با انجام این کار چند نکته مشاهده می شود:

- حداکثر مقدار زمان غیبت بر حسب ساعت، به‌طور قابل‌توجهی از میانگین فاصله دارد و در واقع، این فاصله برابر با ۹ برابر انحراف معیار داده‌ها است! بنابراین این مقدر باید outlier در نظر گرفته شود.
- حداقل مقادیر Reason for absence و Month of absence صفر است! این مقادیر غیرمنتظره در نظر گرفته می شوند زیرا صفر ماه معنی ندارد.
- ستون ID حداکثر مقدار ۳۶ دارد به این معنی که همه این ۷۴۰ رکورد، مشاهدات تکراری هستند.

نگاه دقیق‌تر به ستون‌ها نشان می‌دهد که همه مقادیر صفر در ستون‌های Reason for absence و Month of absence با مقدار صفر در ستون the Absenteeism time in hours مطابقت دارند.
اگر این نتیجه‌گیری نشان‌دهنده هر کارمندی باشد که هیچ روزی غیبت نکرده است، ستون روز هفته نباید پر شود.
بنابراین فقط ردیف‌هایی را نگه می‌داریم که در آن the Absenteeism time in hours==1 باشد.
با انجام این کار تمام اطلاعات مربوط به Disciplinary failure=1 پاک می‌شوند و فقط مقادیر صفر این فیچر باقی می‌ماند؛ بنابراین این ستون برای تحلیل ما سودمند نخواهد بود و آن را drop می‌کنیم.
پس از این کار خواهیم دید که تعداد ID های منحصر به فرد ۳۳ عدد خواهد بود.
با دانستن اینکه ما ۳۳ شناسه منحصربه‌فرد داریم، چارچوب داده‌های موجود را از ۶۹۶ ردیف به ۳۳ ردیف بازآرایی می‌کنیم به صورتی که رکوردهای هر ID با هم ترکیب خواهند شد. این بدان معناست که تمام ردیف‌های دارای ID یکسان به عنوان یک گروه در نظر گرفته می‌شوند.
برای انجام آن:

- مقادیر دسته‌بندی شده Resons for absence به ۲۸ ستون مختلف تبدیل می‌شوند که هر کدام تعداد دفعاتی را که یک ID از یک دلیل خاص استفاده کرده است را نشان می‌دهد.
- همین فرآیند برای ستون‌های Month of Absence، Seasons و Day of the week استفاده می‌شود.
- برای تمام ستون های عددی به جز ستون Absenteeism time in hours میانگین را محاسبه می‌کنیم.
- برای ستون Absenteeism time in hours مجموع این ساعات غیبت محاسبه می‌شود.

با انجام این کار یک DataFrame به نام final_data می‌سازیم تا مجموعه داده‌ای از غیبت کارمندان و معیارهای مربوط به آن را که بر اساس ID گروه‌بندی شده‌اند، آماده کند.

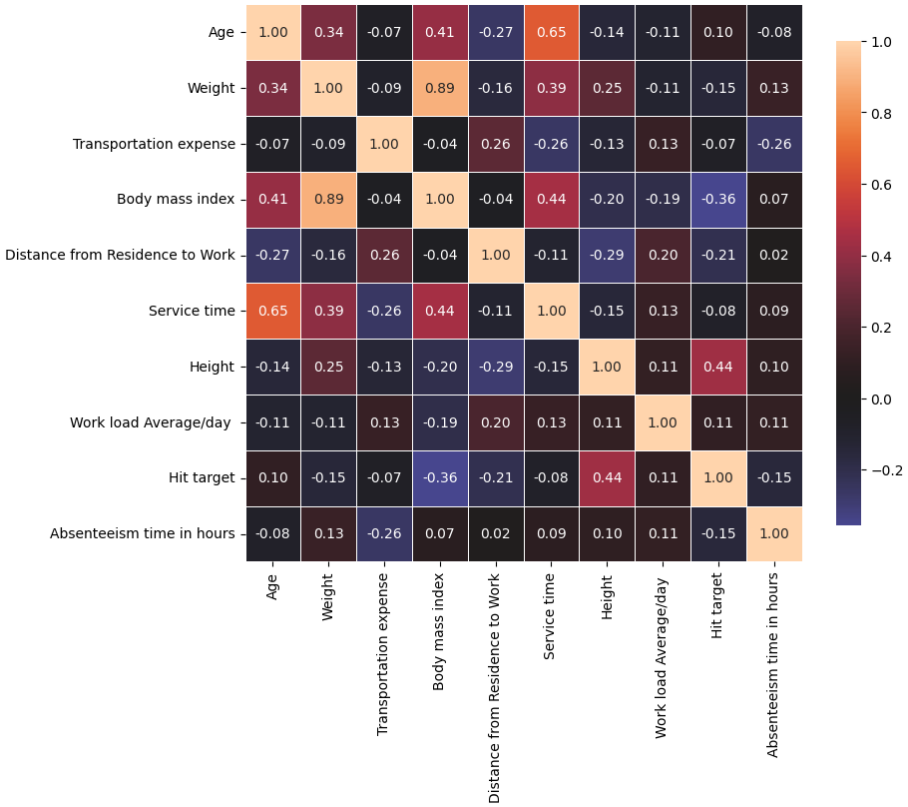
در نهایت این دیتافریم به این صورت خواهد بود:

ID	Absenteeism time in hours	Hit target	Work load Average/day	Age	Transportation expense	Service time	Height	Body mass index	Weight	Social drinker	Social smoker	Distance from Residence to Work	Reason_1	Reason_3	Reason_4
1	121	95.045455	263735.727273	37.0	235.0	14.0	172.0	29.0	88.0	0.0	0.0	11.0	1	0	0
2	25	92.000000	212010.250000	48.0	235.0	12.0	163.0	33.0	88.0	0.0	1.0	29.0	0	0	0
3	482	95.071429	262248.437500	38.0	179.0	18.0	170.0	31.0	89.0	1.0	0.0	51.0	0	0	0
5	104	93.428571	262812.500000	43.0	235.0	13.0	167.0	38.0	106.0	1.0	0.0	20.0	0	0	0
6	72	94.875000	274829.000000	33.0	189.0	13.0	167.0	25.0	69.0	0.0	0.0	29.0	0	0	0

بررسی روابط میان تارگت و فیچرها - Correlation :

برای تجزیه و تحلیل بصری روابط بین ویژگی های مختلف کارکنان از ماتریس correlation آنها استفاده می‌کنیم. این ماتریس با استفاده از heatmap به تصویر کشیده شده است.

همبستگی های قوی (مثبت یا منفی) بین ستون‌ها می تواند ارتباط موثر بین آنها را بیان کند.



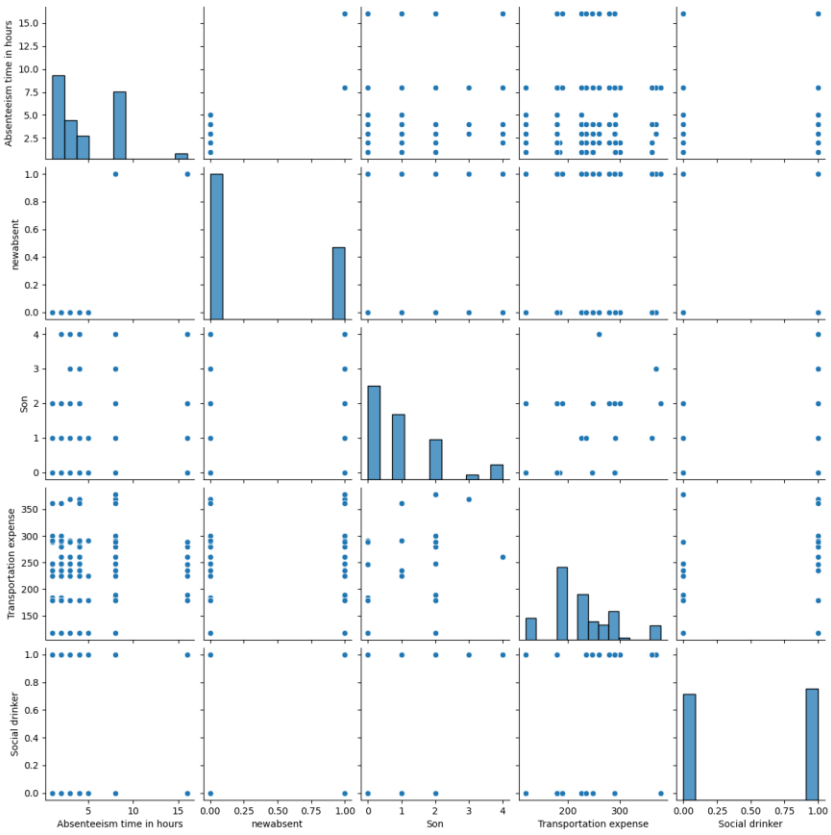
با مقایسه میزان همبستگی ستون Absenteeism time in hours با سایر فیچرها می‌توانیم به میزان مرتبط بودن آنها پی ببریم. همانطور که مشاهده می‌شود میزان همبستگی Age و تارگت ما، منفی است به این معنی که افراد مسن‌تر ساعت غیبت کمتری دارند. همچنین همبستگی میان Weight و تارگت مثبت است و برداشت می‌شود که افرادی که وزن بیشتری دارند، ساعات غیبت آنها نیز بیشتر است.

به دست آوردن Correlation hitmap اختصاص داده شده به متغیر تارگت:

همانطور که مشاهده می‌شود میزان همبستگی تمامی فیچرها با تارگت محاسبه شده و به صورت مرتب شده نمایش داده می‌شود.

```
Absenteeism time in hours          1.000000
newabsent                          0.893610
Son                                0.253280
Transportation expense             0.234481
Social drinker                     0.152384
Social smoker                      0.097577
Pet                                0.073945
Work load Average/day              0.044932
Weight                             0.040171
Body mass index                    0.030359
Month of absence                   0.027230
Height                             0.019867
Age                                0.000715
Seasons                           -0.004881
Distance from Residence to Work   -0.006385
Hit target                        -0.023547
Education                         -0.033962
Service time                      -0.041018
Day of the week                   -0.079943
ID                                -0.139394
Reason for absence                -0.358695
Name: Absenteeism time in hours, dtype: float64
```

نمایش دادن Top Correlation ها :



در این نمودار بیشـــترین میزان همبســـتگی‌ها به نمایش درآمده است.

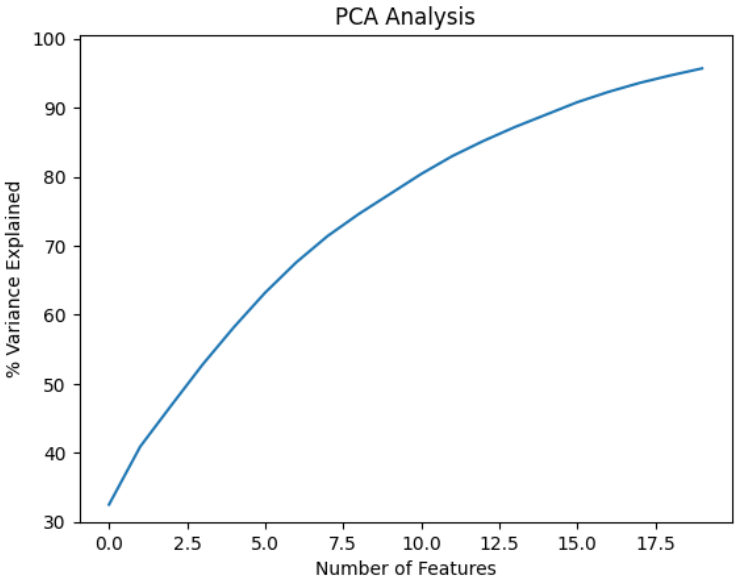
از بیشـترین همبسـتگی ها می‌توان به Absenteeism time in hours و Tranportation expense اشاره کرد.

بررسی روابط میان تارگت و فیچرها - PCA :

در این بخش تجزیه و تحلیل مؤلفه اصلی (PCA) را انجام می دهیم تا حداقل تعداد اجزای اصلی مورد نیاز برای حفظ بیشتر واریانس مجموعه داده را شناسایی کنیم. به عبارت دیگر ابعاد مجموعه داده را کاهش دهیم. تجسم PCA به ما در درک چگونگی توزیع واریانس بین اجزای اصلی مؤلفه های اصلی و اینکه چقدر واریانس خواهند داشت کمک خواهد کرد. علاوه بر این می‌توانیم ویژگی هایی را انتخاب کنیم که بیشترین اهمیت را دارند، نویز را کاهش و راندمان محاسباتی را بهبود ببخشیم.

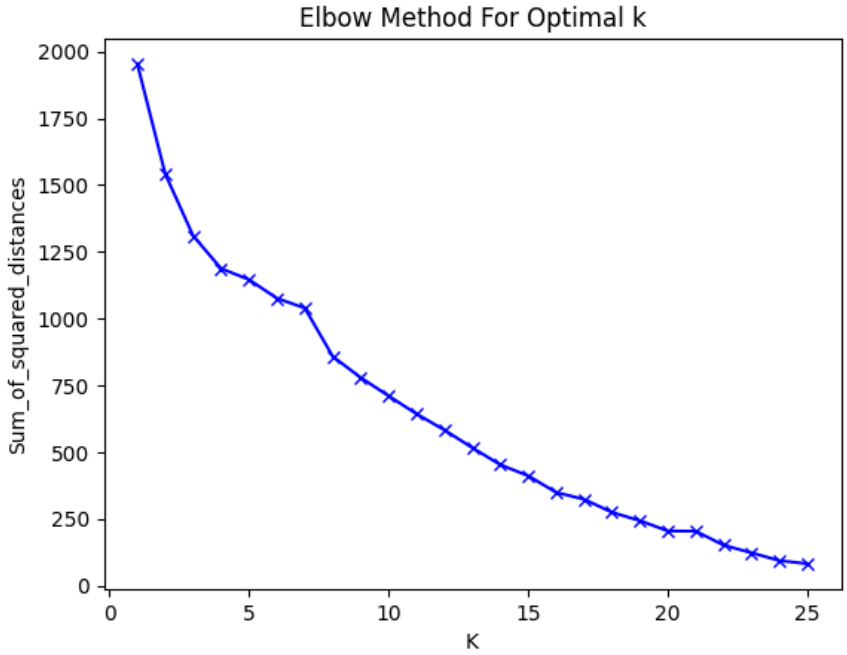
در دیتافریم ما، ۶۲ ستون وجود دارد، استفاده از PCA قبل از خوشه بندی، کاهش متغیرها و استفاده از اجزای غیرهمبسته را تضمین می کند. نمودار حاصل از این کد Elbow Point را نشان می دهد. این نقطه جایی است که اجزای اصلی اضافی کمتر به واریانس کمک می کنند.

شکل زیر نشان می دهد که ۲۰ مؤلفه اصلی ۹۵ درصد از واریانس کل را توضیح می دهند.



بررسی روابط میان تارگت و فیچرها - Clustering :

حال به سراغ خوشه‌بندی می‌رویم. برای تعیین تعداد بهینه خوشه ها (k) و گروه بندی داده ها با استفاده از روش Elbow از ترکیب تجزیه و تحلیل مؤلفه اصلی (PCA) و خوشه بندی K-Means استفاده کردیم. K بهینه در elbow point در نمودار مشخص می شود که بر اساس داده ما مقداری بین ۰ تا ۲۳ است.



می بینیم که از این نتیجه بالا، باید از ۱۵ تا ۲۰ خوشه استفاده کنیم و از آنجایی که ما فقط ۳۳ شناسه مشتری داریم، می‌توانیم ببینیم که روش خوشه‌بندی نمی‌تواند کارمندان گروه‌های متمایز را پیدا کند.

اگرچه، این روش با نگاه کردن به هر شناسه منحصربه‌فرد بی‌نتیجه است، داشتن داده‌های بیشتر با کارمندان بیشتر به ما بینش عمیقی می‌دهد.

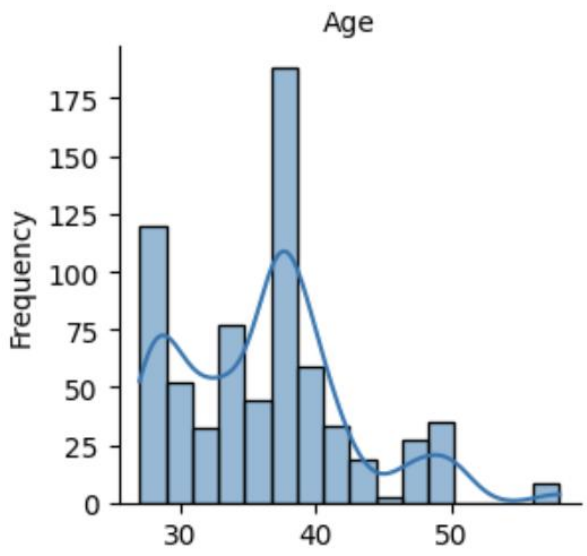
راهبرد به کار گرفته شده در تحلیل اصلی: مشاهده شد که با مجموعه دیتای ما، نتیجه مطلوبی حاصل نمی‌شود. بنابراین به همان دیتای اولیه با ID ۷۴۰ برمی‌گردیم.

حال نگاه دقیق‌تری به هر یک از ستون‌ها می‌اندازیم.

برای درک بهتری از توزیع دیتا، نمودار توزیع هر یک از فیچرها را رسم می‌کنیم.

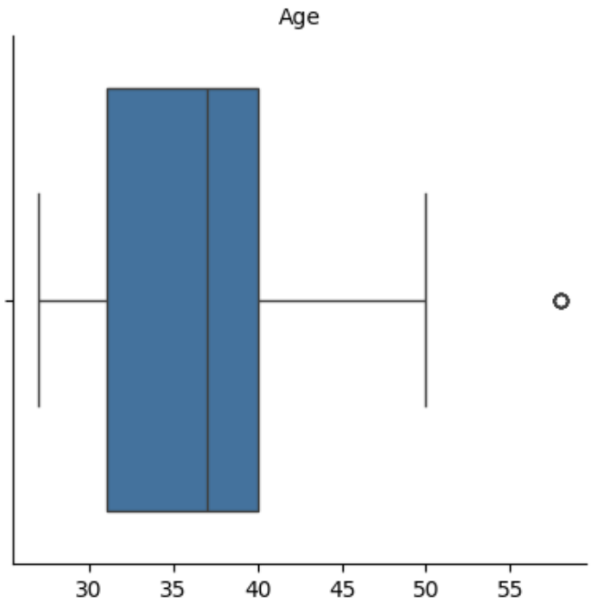
مثال - بررسی ستون Age:

با بررسی این ستون می‌توانیم به این سوال پاسخ دهیم که کارمندان این شرکت در چه رنج سنی هستند؟ اغلب جوان هستند یا میانسال و ... .



همانطور که مشاهده می‌شود اغلب کارمندان در اواخر دهه ۲۰ و ابتدای ۴۰ سالگی هستند.

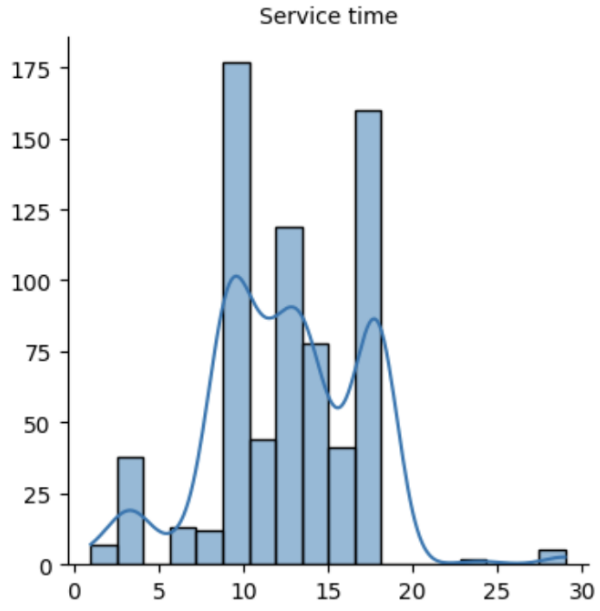
نمودار boxplot فیچر Age:

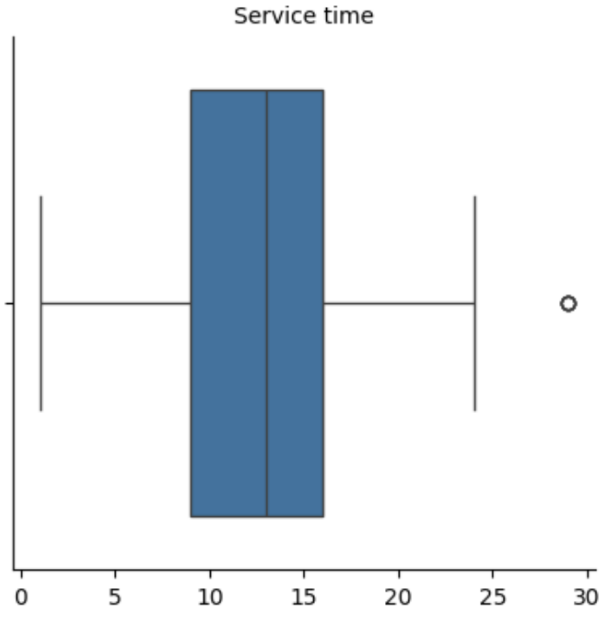


با استفاده از این نمودار نیز می‌توانیم دید بهتری از توزیع‌شدگی دیتا داشته باشیم.

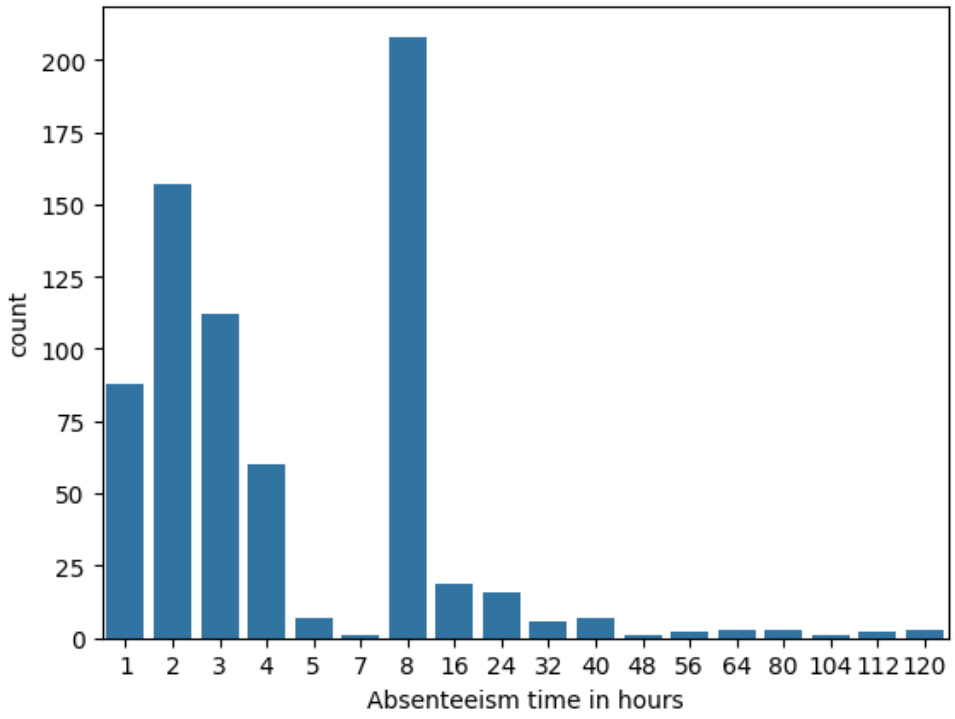
همانطور که مشاهده می‌شود سنین بالای ۵۵ به عنوان outlier در نظر گرفته می‌شوند و با حذف آن می‌توان آنالیز بهتری انجام داد.

مثال - بررسی ستون Service time:





مثال - بررسی ستون Absenteeism time in hours:



با بررسی این مثال‌ها در می‌یابیم که بیشتر کارگران در اواخر دهه ۲۰ تا اوایل دهه ۴۰ زندگی می کنند. علاوه بر این، بیشتر کارمندان بین ۱۰ تا ۱۸ سال سابقه خدمت دارند.

همچنین در ستون زمان غیبت در ساعت نکات زیر به چشم می‌خورد:

- طرح بیشتر به سمت راست که بیشتر مقادیر کمتر از ۸۰ ساعت غیبت هستند، منحرف است.
- در واقع ۶۱٪ کارمندان کمتر از ۸ ساعت و ۲۹٪ آنها دقیقاً ۸ ساعت غیبت داشته اند.
- بنابراین ۹۱٪ زمان غیبت بر حسب ساعت با ۸ ساعت یا کمتر غیبت توضیح داده می شود.

با توجه به مقادیر خاص آن که به وضوح پیوسته نیستند، استنباط می کنیم که این مقادیر categorical هستند به این معنی که باید از Classification برای مدل سازی استفاده کنیم که جزو موارد مجاز به استفاده در این پروژه نمی‌باشد.