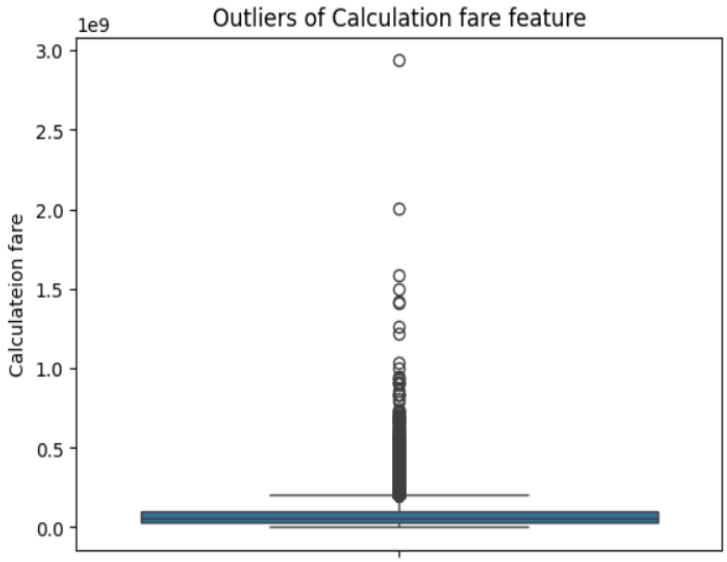
به نام خدا

ادامه توضیحات فایل نوت بوک کولب : Transportation_Dataset1

تحلیل نمودار رو به رو :



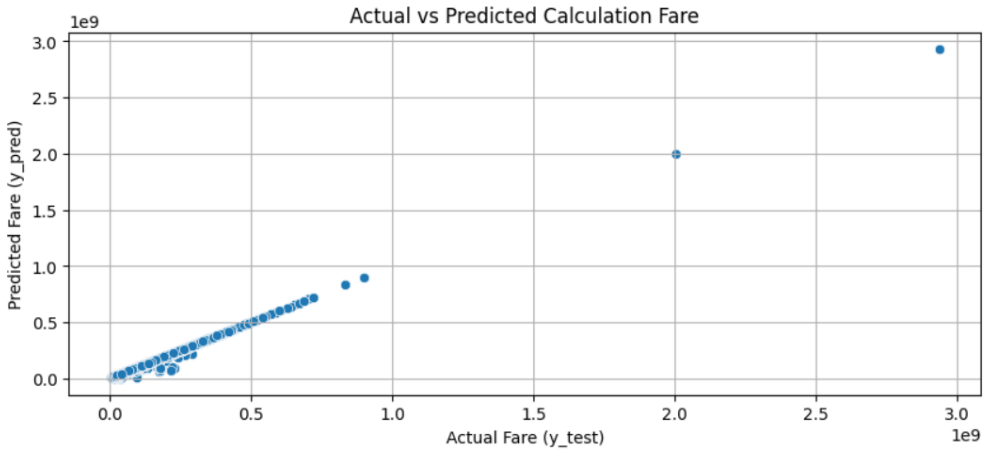
با توجه به نمودار بالا می توان دریافت که حجم بیشتر مبالغ کرایه محاسباتی بین ۰ تا ۲۵ میلیون است . همینطور می توان دریافت که داده های پرت از ۲۵ میلیون تا ۳۰۰ میلیون تومان هستند.

این نمودار با استفاده از دستور sns.scatterplot() رسم شده است.

الگوریتم Linear Regression

با توجه به نمودارهای زیر تحلیل های خود را انجام می دهیم :

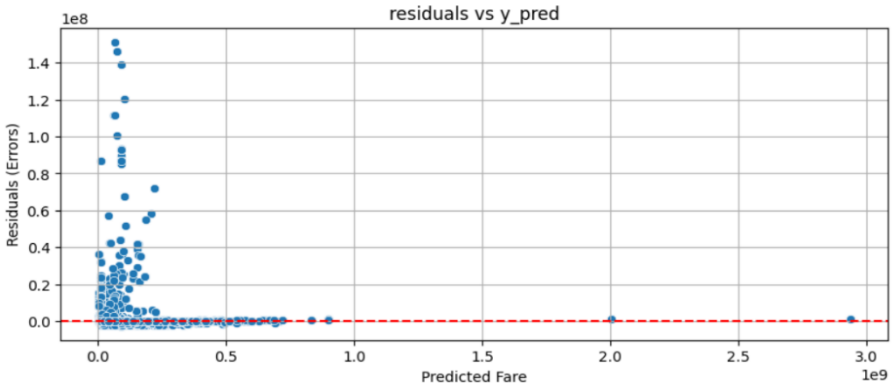
با توجه به نمودار Actual vs Predicted plot (نمودار بیش بینی در برابر مقدار واقعی) نقاط بر روی نمودار نشان می دهند که مقدار واقعی با مقدار پیش بینی شده ی کرایه محاسباتی همخوانی دارند.



اکثریت نقاط در نزدیکی خط مورب هستند و این یعنی پیش بینی ها قابل قبول اند. اما چند نقاط بسیار پرت (Outlier) هم در نمودار دیده می شوند که مدل آنها را بد پیش بینی کرده است. پس به احتمال زیاد این داده ها ، داده های پرت هستند.

در واقع ، مدل به طور کلی عملکرد خوبی داشته است ولی با توجه به اینکه تاثیر داده های ناهنجار بالاست ، یعنی چند رکورد از کرایه ها مبالغشان بسیار بالا بوده و همین باعث کاهش دقت و کشیده شدن نمودار شده است.

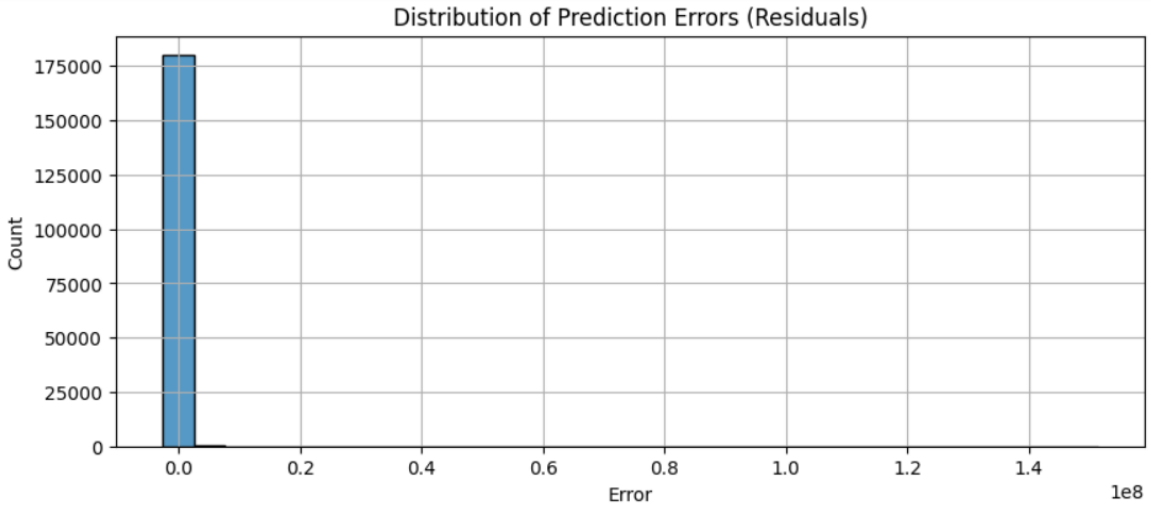
نمودار روبه رو (نمودار خطاها نسبت به پیش بینی) نشان دهنده باقی مانده ها (اختلاف مقدار و



پیش بینی شده) در برابر مقادیر پیش بینی شده رسم شده اند. همانطور که می توان دید ، بیشتر خطاها تطراف صفر هستند یعنی این مدل در بیشتر پیش بینی ها نزدیک به مقدار درست است. اما چند خطای بسیار زیاد (در مقیاس 8^10) داریم ؛ این نقاط می توانند رکوردهایی با کرایه محاسباتی بسیار بالا و یا اشتباه باشند.

تحلیل : توزیع خطا در نمودار متقارن نیست یعنی مدل روی برخی نقلط ناتوان بوده ، شاید چون داده ها اسکلیت نرمال ندارند و یا ناهنجاری در انها وجود دارند.

نمودار زیر (نمودار توزیع خطاها) نشان می دهد که یکسری از پیش بینی ها چه مقدار خطا دارند.



در نمودار مشاهده می شود که اکثرا خطای نزدیک صفر دارند و به این معناست که مدل عملکرد خوبی داشته است. ولی یک دنباله بلند داریم و این یعنی تعداد کمی از رکوردها دارای خطای خیلی زیادی هستند.

تحلیل : مدل در اکثر مواقع دقیق است ، اما چند رکورد با خطای شدید وجود دارد که توزیع را کمی خراب کرده است.

بررسی اعداد ارزیابی مدل

معیار	مقدار	تحلیل
MAE	128575.50699209691	این مقدار مناسب است ، اما به شدت وابسته به مقیاس داده است.
MSE	3446518422398.2856	این عدد بسیار بزرگ است ؛ نشان دهنده این است که مقادیر پرت خطا بر روی خطا تاثیر منفی گذاشته اند.
R2 Score	0.9990293597437907	این عدد بسیار عالی است یعنی مدل ۹۹.۹٪ از واریانس داده ها را توضیح داده است.

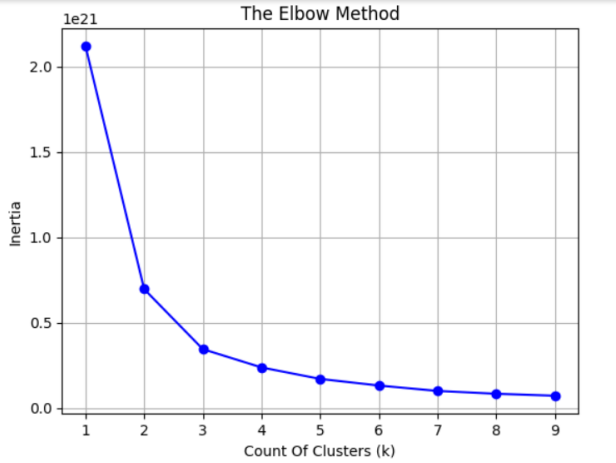
تحلیل کلی جدول بالا : با وجود اینکه R2 بسیار خوب است ، اما مقدار زیاد MSE نشان می دهد که چند داده ی شدیدا پرت وجود دارند که مدل نتوانسته آنها را به درستی یاد بگیرد.

می توانیم برای داده های پرت از روش نرمال سازی استفاده کنیم .

می توانیم در ادامه از مدل های مقاوم در برابر داده های پرت مثل Hubber Regression یا Random Forest Regressor استفاده کنیم زیرا این مدل ها در برابر داده های پرت حساس نیستند.

می توانیم داده های پرت را شناسایی و حذف کنیم ؛ می توانیم از روش های IQR یا Isolation Forest استفاده کنیم.

مدل KMeans Clustering با Elbow



وقتی ٪inertia (مجموع فاصله هر نقطه تا مرکز هر خوشه) را بر حسب تعداد خوشه‌ها رسم کردیم، انتظار داریم در نقطه‌ای کاهش تند و از آن‌جا به بعد تغییر کمتر باشد. روش Elbow به ما کمک می کند بفهمیم بهترین K کدام است.

در نمودار روبه رو Elbow در حوالی k=3 و یا k=4 قرار دارد. یعنی افزایش خوشه بیش از این تعداد k باعث کاهش inertia نمی شود. در کد KMeans Clustering ، k برابر با ۴ انتخاب شده است.

در کد مربوط به KMeans Clustering داده های کرایه محاسباتی به ۴ گروه تقریبا هم رده دسته بندی شده اند. مثل : . گروه ۰ (کرایه های پایین) ، گروه ۱ (کرایه های متوسط) ، گروه ۲ (کرایه های بالا) ، گروه ۳ (کرایه های خیلی بالا).

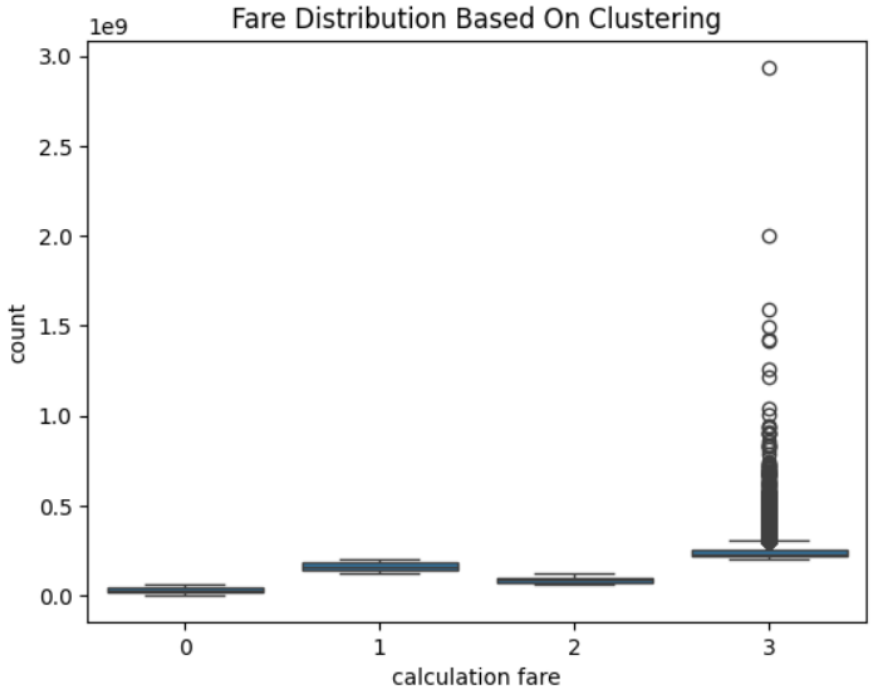
انجام این خوشه بندی به ما کمک می کند تحلیل کنیم چه بخشی از سفارش ها خیلی ارزان یا خیلی گران هستند.

می توانیم با یک خط کد transport_df['kmeans_cluster_cal_fare'].value_counts() تعداد مقادیر هر خوشه را مشاهده کنیم.

تحلیل : اگر خوشه گروه ۳ (کرایه های خیلی بالا) خیلی کم باشد ، یعنی فقط چند سفارش با کرایه های بالا وجود دارد و این می تواند پرچم تقلب ، اشتباه ورودی و یا خدمات خاص باشد.

نمودار زیر ، نمودار توزیع خوشه هاست که در ادامه به بررسی و تحلیل ان می پردازیم. موقعیت خوشه‌ها در محور افقی همانطور که در بالا توضیح داده شد ، بخش های ۰ تا ۳ نمایش دهنده ۴

خوشه هستند. گروه ۰ (کرایه های پایین) ، گروه ۱ (کرایه های متوسط) ، گروه ۲ (کرایه های بالا) ، گروه ۳ (کرایه های خیلی بالا).



مرکز داده ها (میانه و خط میان چارکی)

میانه خوشه ۰ در پایین ترین سطح قرار گرفته است و نمایانگر کرایه ها پایین است.

میانه خوشه ۳ بسیار بالاتر از بقیه است ، که نشان می دهد که کرایه ها خیلی بالا هستند.

خوشه های میانی (۱ و ۲) با یک فاصله قابل مشهود از هم قرار گرفته اند و این یعنی تقسیم بندی منظم صورت گرفته است.

گستره (IQR) و پراکندگی

خوشه ۰ و ۱: IQR بزرگ‌تری دارند، نشان‌دهنده تنوع در مقادیر کرایه (گستره وسیع‌تری از قیمت‌ها) است.

خوشه ۲ و ۳: IQR باریک‌تر، یعنی داده‌های نسبتاً متمرکزتر در محدوده خاص‌شان هستند.

این معنی را می‌دهد که در خوشه‌های میانی و بالا، قیمت‌ها نسبتاً نزدیک هم هستند اما هنوز کاملاً یکنواخت نیستند.