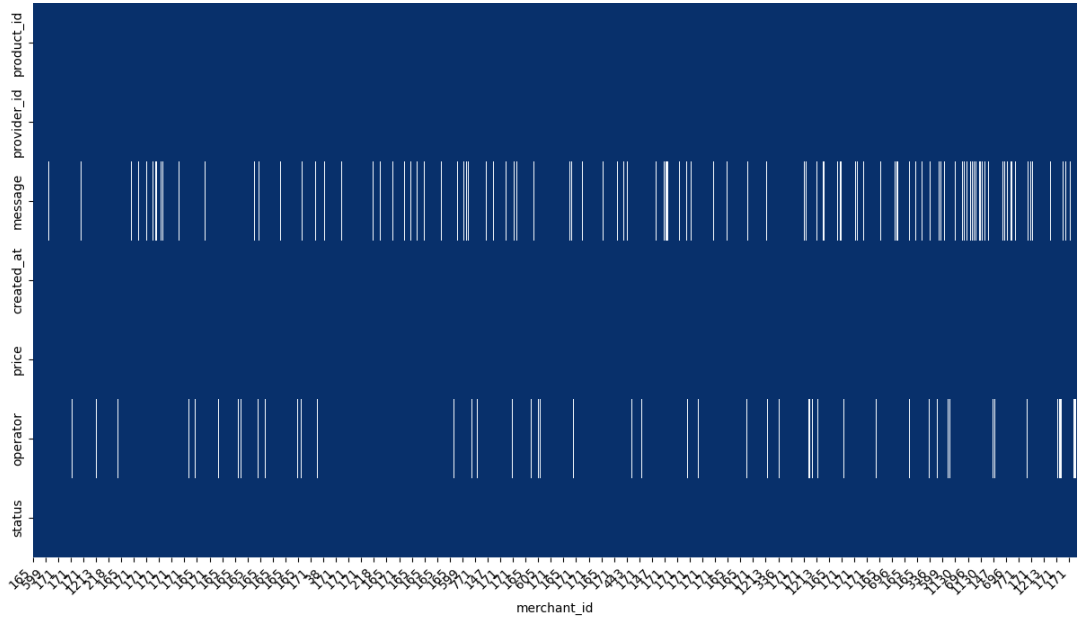
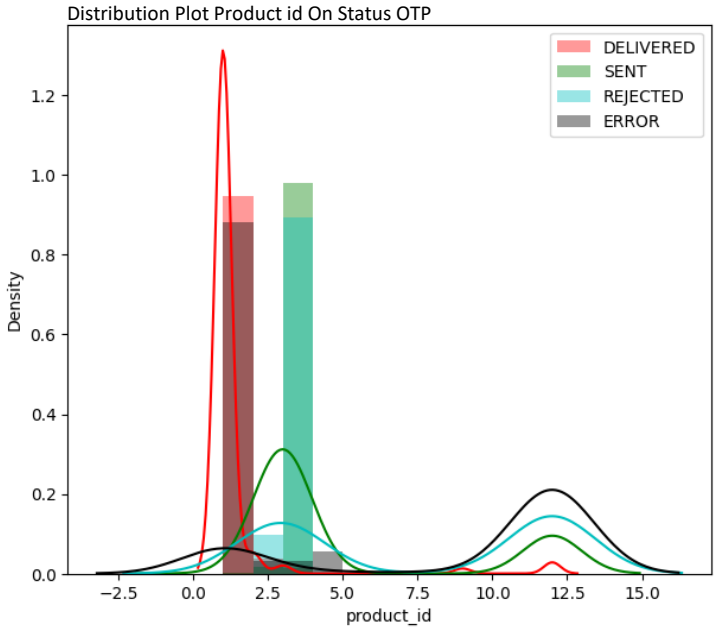
Report: Classification with EDA

Missing Value Plot



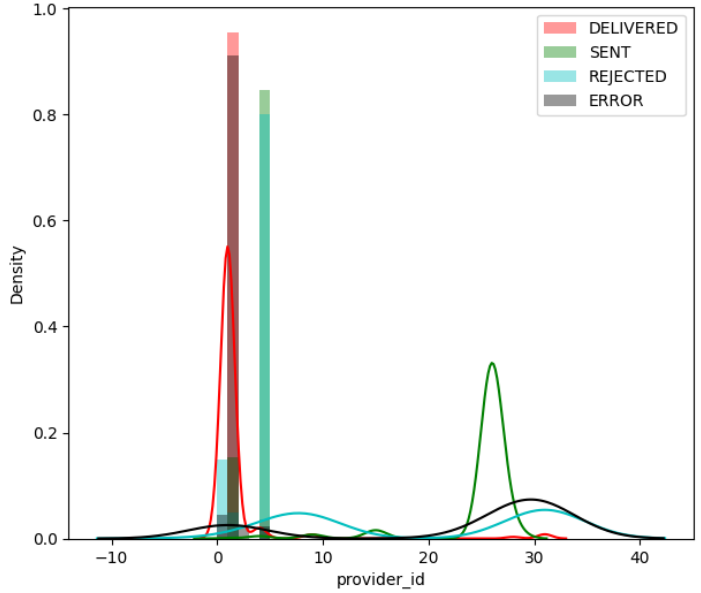
Distribution Plot Product id On Status OTP



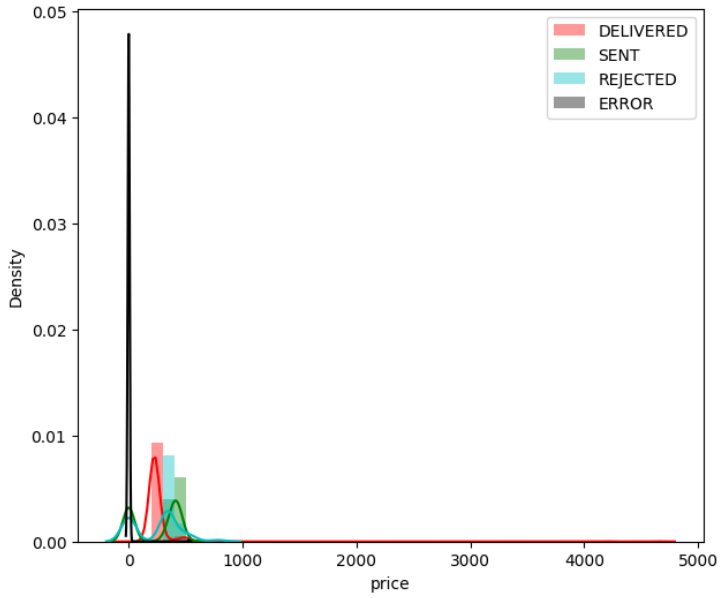
product_id	1	2	3	4	5	6	7	9	12
status									
DELIVERED	7745	299	118	0	0	0	0	74	165
ERROR	81	3	3	5	0	1	2	4	281
REJECTED	0	14	128	0	1	0	0	3	158
SENT	2	11	690	0	0	0	0	0	212

provider_id	0	1	2	4	5	8	9	15	16	17	23	26	28	31
status														
DELIVERED	74	7760	0	284	0	15	103	0	3	6	0	0	43	113
ERROR	4	83	2	2	0	0	0	0	0	0	5	4	120	160
REJECTED	3	1	0	15	1	128	0	0	0	0	0	0	0	156
SENT	0	2	0	11	0	0	18	38	0	0	0	796	50	0

Distribution Plot Provider id On Status OTP



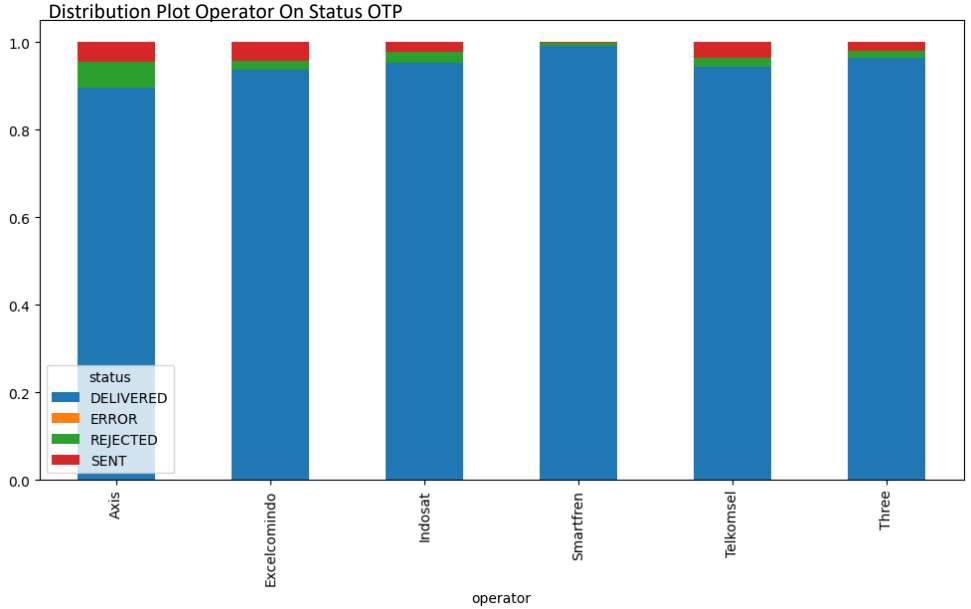
Distribution Plot Price id On Status OTP



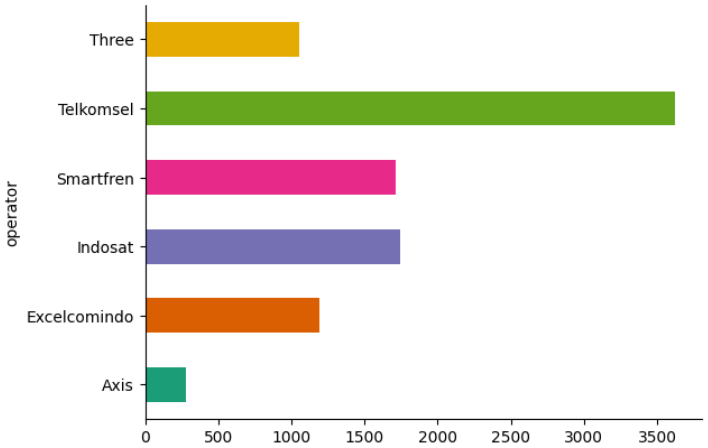
price	Cheap	Expensive	Vvip
status			
DELIVERED	8289	65	35
ERROR	0	1	0
REJECTED	175	18	0
SENT	536	0	0

- Missing value plot: to find out the contribution of values to the data frame. several columns are dropped, such as id, phone and trx_id. with consideration of using the missing value plot above, I had to take the initiative to drop due to the lack of maximization of the data set. then the data set with the new data frame as above can be processed.
- Distribution Plot Product id On Status OTP: The above plot contains information about delivered, sent, rejected, and errors.and Product_id has a very high reference to the status in numbers 1, 2, 3 And 12, seen from the crossover table shows that product_id has a very high DELIVERED value. but has the highest ERROR value also compared to other components.
- Product_id is indicated to have a very high value on the status in numbers 1, 4, 9 and 31, it can be seen from the crossover table that product_id has a very high DELIVERED value. but number 31 also has the highest error value compared to other components. then the important thing is also in number 8 which has a fairly high REJECTED number and number 26 which has a fairly high SENT value as well.
- Price has quite a lot of data variants, therefore I took the initiative to binning the data into categorical data with parameters 0-500 is cheap, 500-1000 is expensive and 1000 onwards is vvip.
- price with cheap values tends to have a very high value and the value of delivered content is very high too. with data like this, it can be concluded that the segmentation of customers is with a low price value.

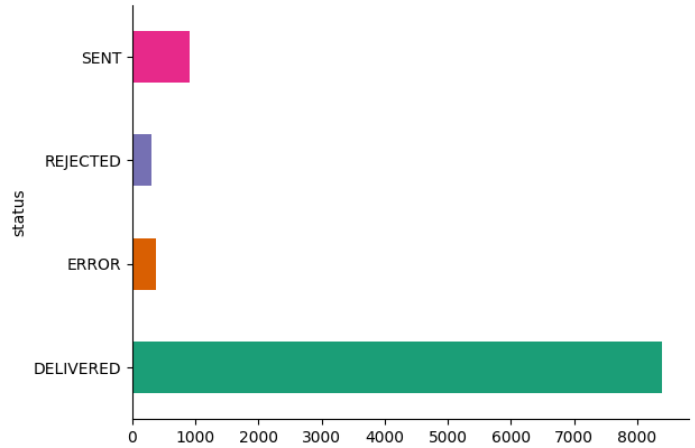
Distribution Plot Operator On Status OTP



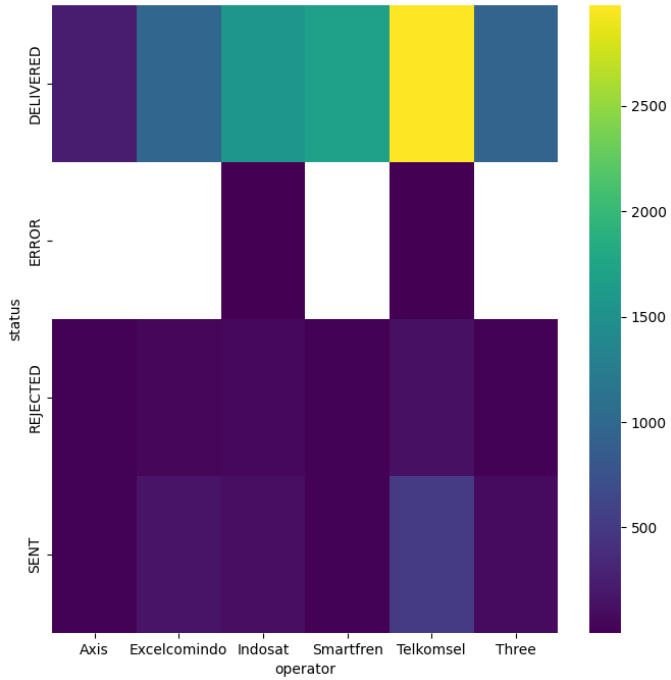
Operator Barplot



Status Barplot



Operator vs Status Heatplot



Preprosesor Scor Result

```
 0.99962505 0.99987502        nan        nan 0.99962505 0.99987502
 0.99937498 0.99974998        nan        nan 0.99950002 0.99987502
 0.99825008 0.99974998        nan        nan 0.99825008 0.99974998
 0.99825008 0.99974998        nan        nan 0.99837511 0.99987502
 0.99937498 0.99974998        nan        nan 0.99937498 0.99974998
 0.99937503 0.99987502        nan        nan 0.99937503 0.99987502
 0.999625   0.99987502        nan        nan 0.999625   0.99987502
 0.99850005 0.99962495        nan        nan 0.99850005 0.99962495
 0.99850005 0.99962495        nan        nan 0.99850005 0.99962495
 0.99862508 0.99962495        nan        nan 0.99862508 0.99962495
 0.99825003 0.99962495        nan        nan 0.99825003 0.99949997
 0.99800011 0.99962495        nan        nan 0.99800011 0.99949997
 0.99800011 0.99962495        nan        nan 0.99800011 0.99949997
 0.99787508 0.99949997        nan        nan 0.99787508 0.99949997]
  warnings.warn(
{'algo__n_neighbors': 3, 'algo__p': 1, 'algo__weights': 'distance'}
1.0 0.9998750156230471 1.0
```

- Distribution of Carrier plots by Status: The majority of messages across all providers are in the DELIVERED status (blue). A small percentage of messages are in the REJECTED (green) and SENT (red) status. The REJECTED (yellow) status is almost non-existent or very minimal for all providers. Consistency All providers show similar patterns of message status distribution. This chart can be used to assess the performance of service providers in terms of message delivery success rates.
- the operators plotted above: Telkomsel is the leading operator in this graph, with the highest value. axis appears to be the smallest among operators based on the values represented.Smartfren and Indosat have similar values, indicating comparable performance or metrics in this context.
- the operators plotted above: DELIVERED being the highest status in this graph, with the highest value.REJECTED seems to be the smallest among the other states. based on the values represented. Therefore, through this data, it can be concluded that the performance of the provider is quite good.
- Preprocessing is a technique used to transform raw data in a useful and efficient format. This initiative is necessary because raw data is often incomplete and has an inconsistent format. Data quality itself has a direct correlation with the success of any project that involves data analysis. there is an option to use data imputation techniques and here I used minmax scaling imputation and for training data I used a combination of algorithms from jcopml and scikit learn with parameters using gsp parameters. which resulted in a score of 99%.

Feature Engenering

- This preprocessor is used as an initial benchmark because a score of 99% raises deeper questions as to whether this data is unbalanced or overhit.
- The next homework is to try scaling using RobustScaler and Normalizer and try with other parameters. Clustering using K-modes, K-means and K-prototype. and using prediction models.