CATEGORICAL CORRELATIONAL ANALYSIS ON FREIGHT SPOT RATES

Progress Report

CONTENTS OF THE REPORT

AUTHORS OF THE REPORT

Prakhar Gupta
Supervisor

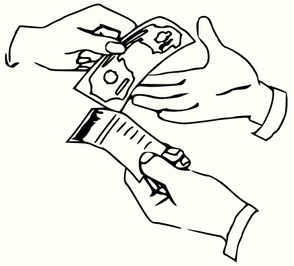
Amartya Chaudhury
Supervisor

Sparsh Amarnani
Data Science Intern

SECTION 01:

INTRODUCTION AND DATA WRANGLING

INTRODUCTION



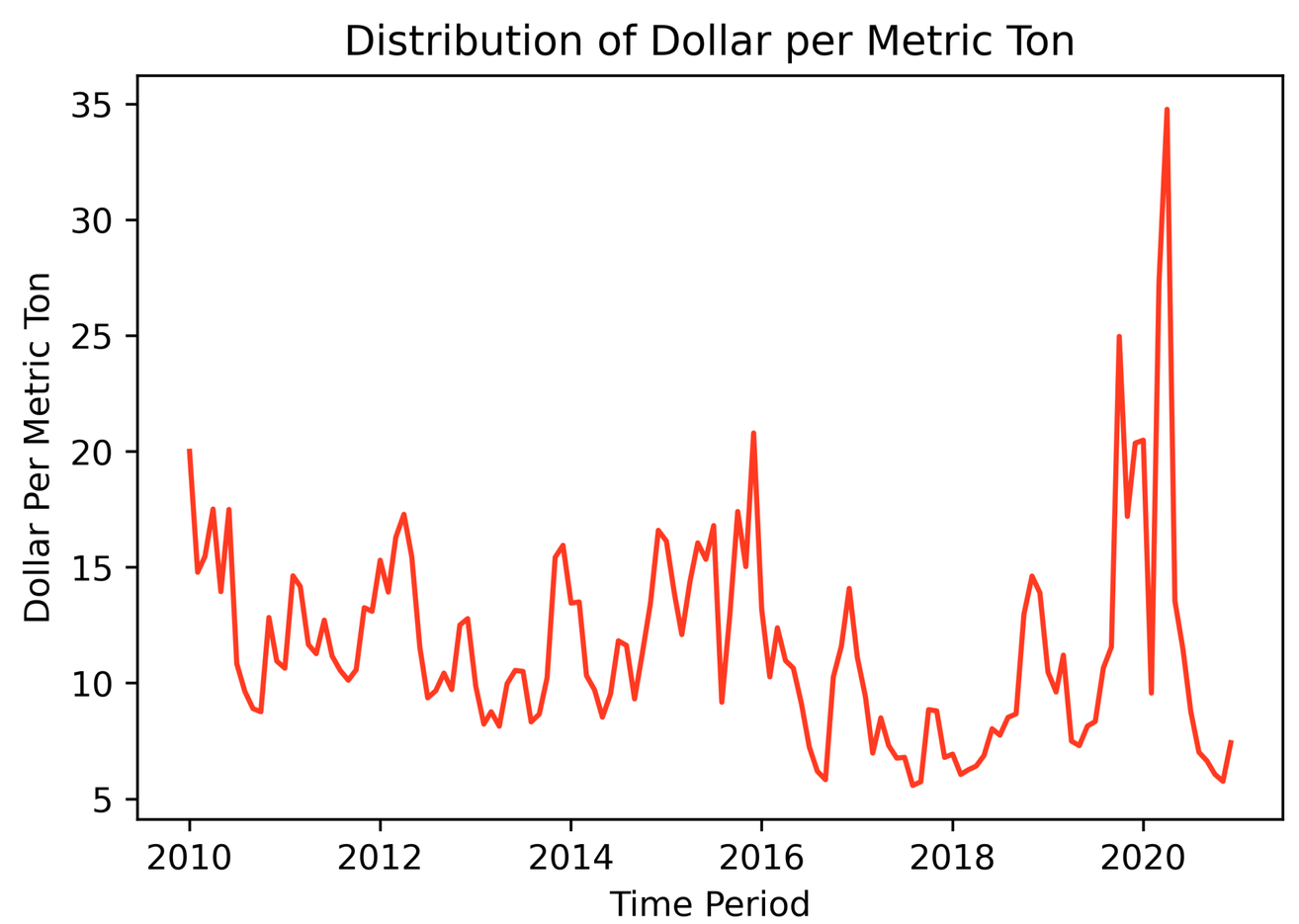
The freight market has two types of transactions. The first one is the time charter under which the ship is hired by the day. The second one is the freight contract in which the shipper buys transport from the shipowner at a fixed price per ton of cargo. Our presentation is centered on the second transaction.

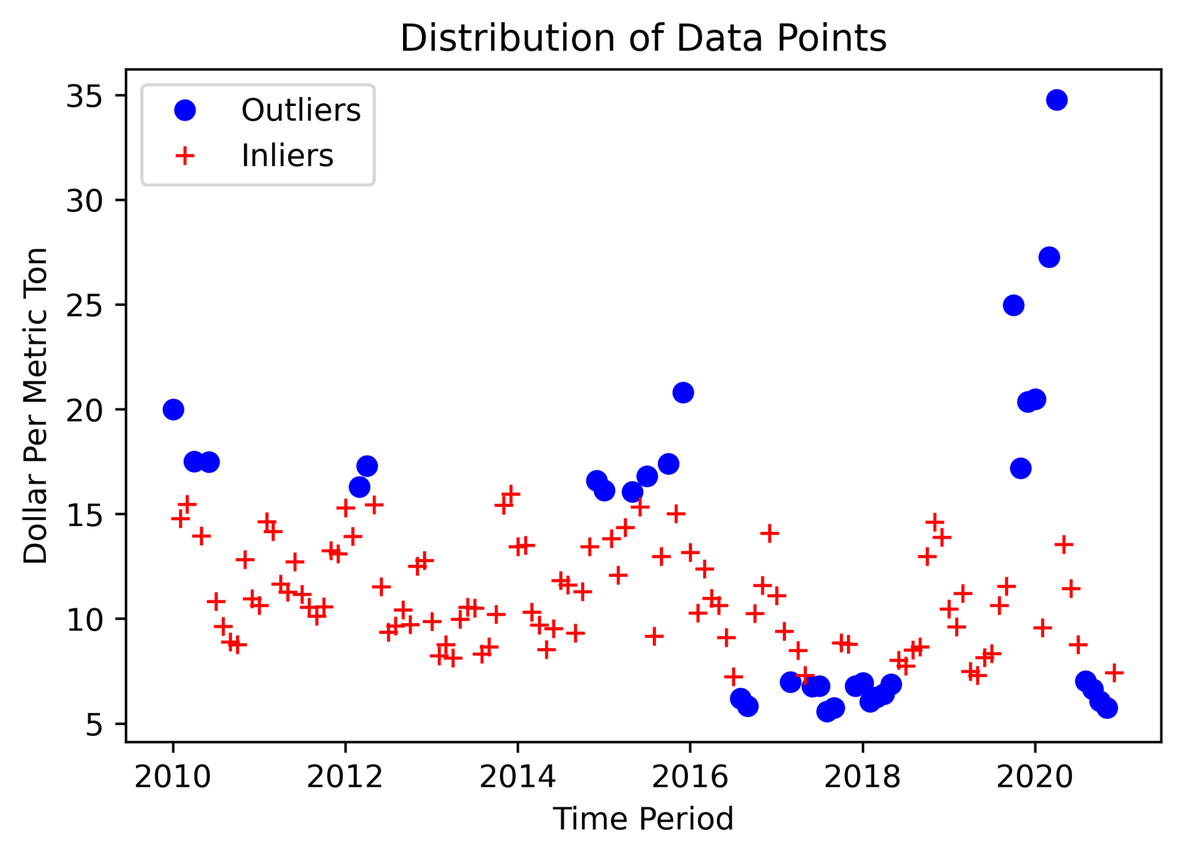
We wish to identify outlying values of Dollar per Metric Ton which is our target variable and highlight features that correlate with our inlying and outlying target variables values with and without time lag to find landmark parameters.

Our data is a monthly time series ranging 10 years from 2010 to 2020 and having 108 parameters. It tracks BDTI, flat rates, dollars per metric ton and other data points for the TD3 route.

DEFINING AND CREATING OUTLIERS

Our goal is to find a methodology which will include the 2020 spike in freight rates as outliers and define other values as outliers too.



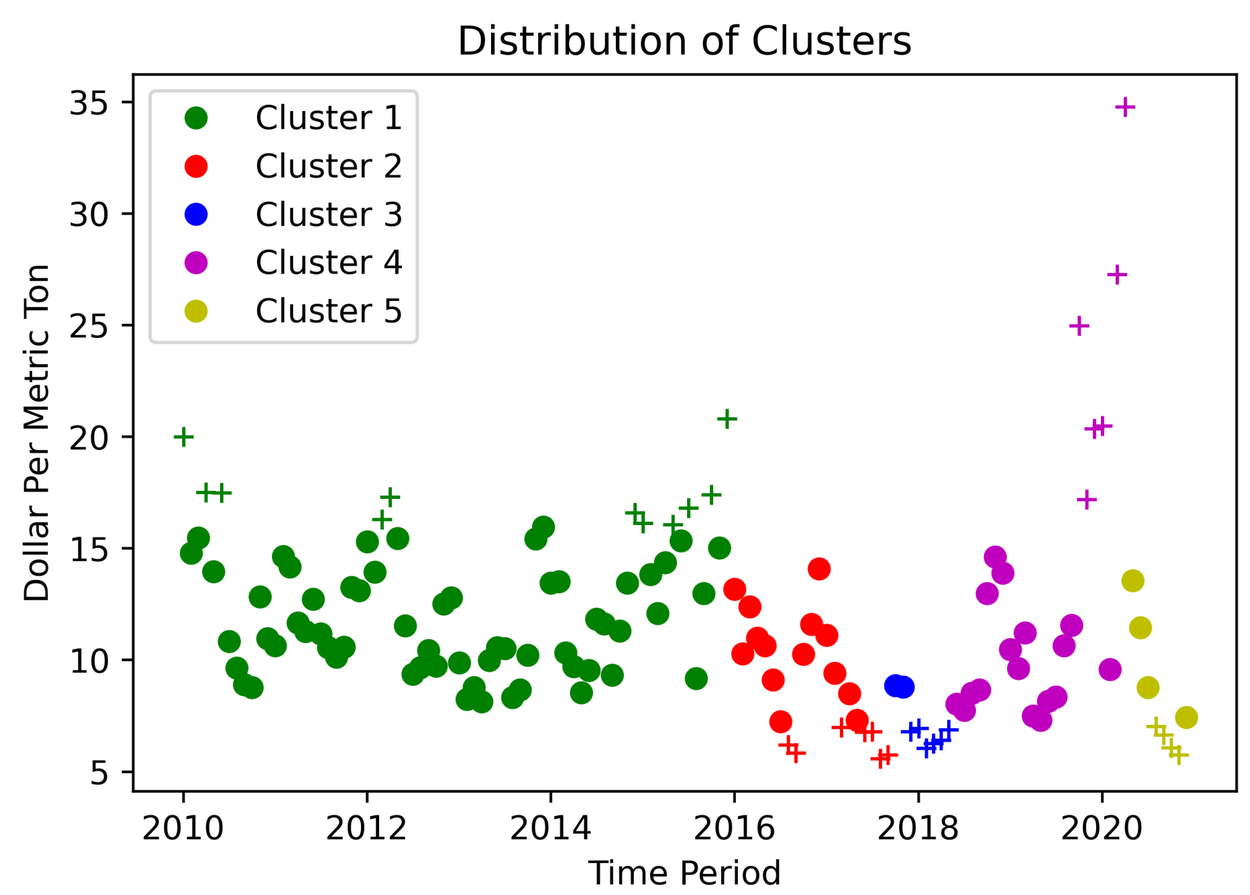


We calculated the Zscore values of our target variable with mean and standard deviation and picked the values with a Zscore higher than 1 or lower than -1 as Outliers.

We found 34 target values which were then grouped according to their closeness in value and time series.

ESTABLISHING CLUSTERS OF OUTLIERS

- The outliers were analyzed for our clusters based on proximity in time series and values. 5 clusters were found.
- The inliers were then grouped into the same clusters based on the inclusive time period.
- In all clusters the outlying and inlying clusters had distinct ranges.



ADDING TIMELAG TO DATA

Exception	Dollar_pe	BDTI TD3	Flat Rate	Earnings	UL/VLCC C	UL/VLCC C	VLCC Orde	VLCC Flee
0.00	14.79067	79.01	18.72	39970.65	198	61549482	38.34	535
0.00	15.47582	82.67	18.72	43224.09	193	60058122	37.6	531
1.00	17.51818	93.58	18.72	53740.25	189	58806458	36.978	528
0.00	13.95576	74.55	18.72	35195.68	184	57297704	35.803	531
1.00	17.50133	93.49	18.72	56926.95	186	57973199	36.128	532
0.00	10.83701	57.89	18.72	19528.68	185	57700265	35.88	533
0.00	9.642672	51.51	18.72	12233.05	187	58354811	36.132	535
0.00	8.905104	47.57	18.72	8757.64	185	57749405	35.525	538
0.00	8.768448	46.84	18.72	5663.62	191	59708510	36.582	540
0.00	12.83818	68.58	18.72	26923.09	197	61633089	37.818	539
0.00	10.95682	58.53	18.72	15053.67	200	62602956	38.24	541
0.00	10.64931	47.1	22.61	10432.45	192	60113666	36.571	543
0.00	14.63771	64.74	22.61	26867.05	192	60124281	36.495	544

Time lag : Is a delay that we can add to a time series data to align past data values of certain other attributes to present data values of other attributes. This was done to find correlation between lagged attributes of our dataset with non lagged Dollar per Metric Ton and to find causation between lagged data points of our dataset to predict future exceptions.

We created four time lags: 1 month, 2 month, 3 month, 4 month.

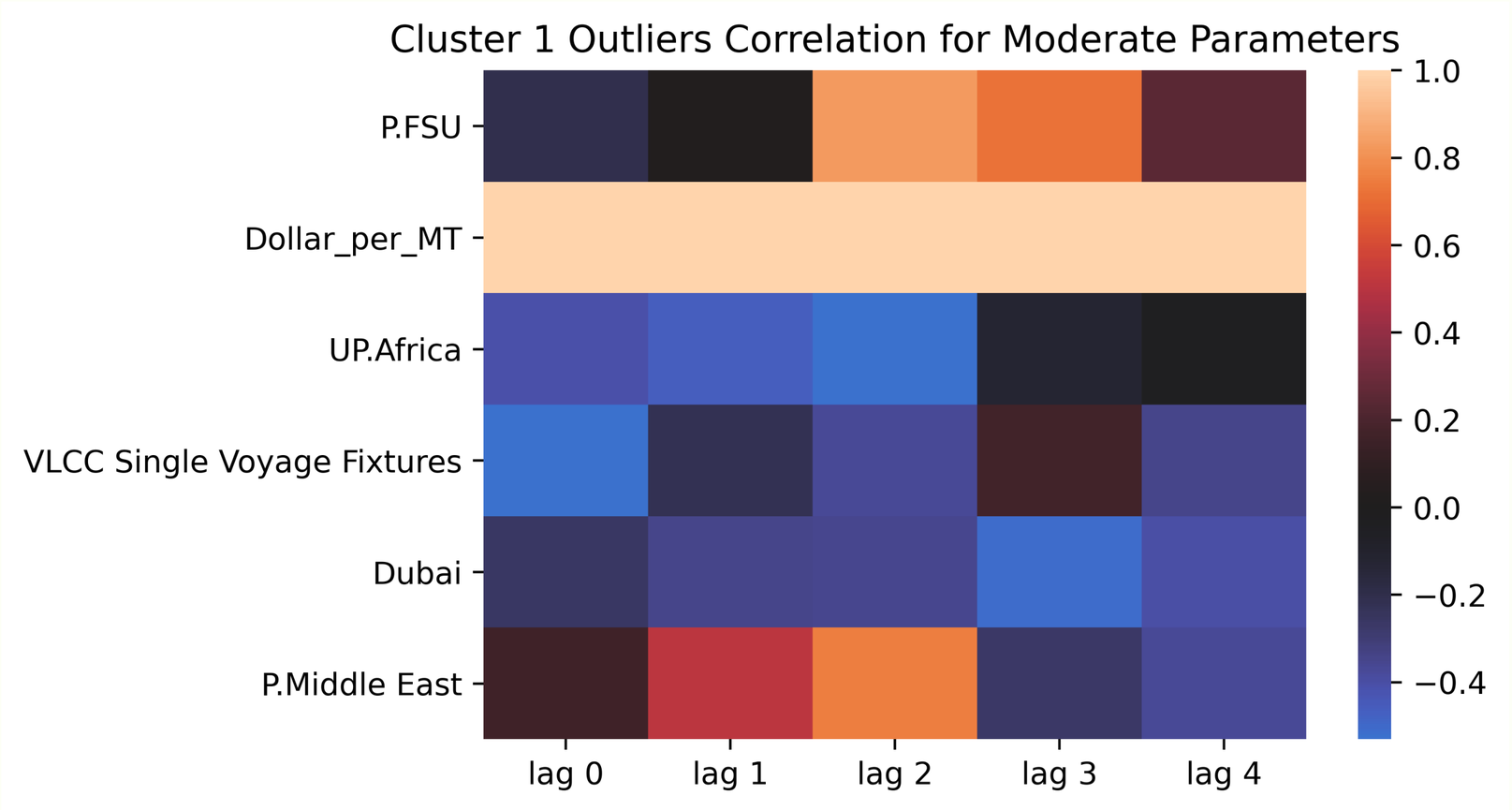
SECTION 02:

ANALYSIS DONE ON OUR DATA

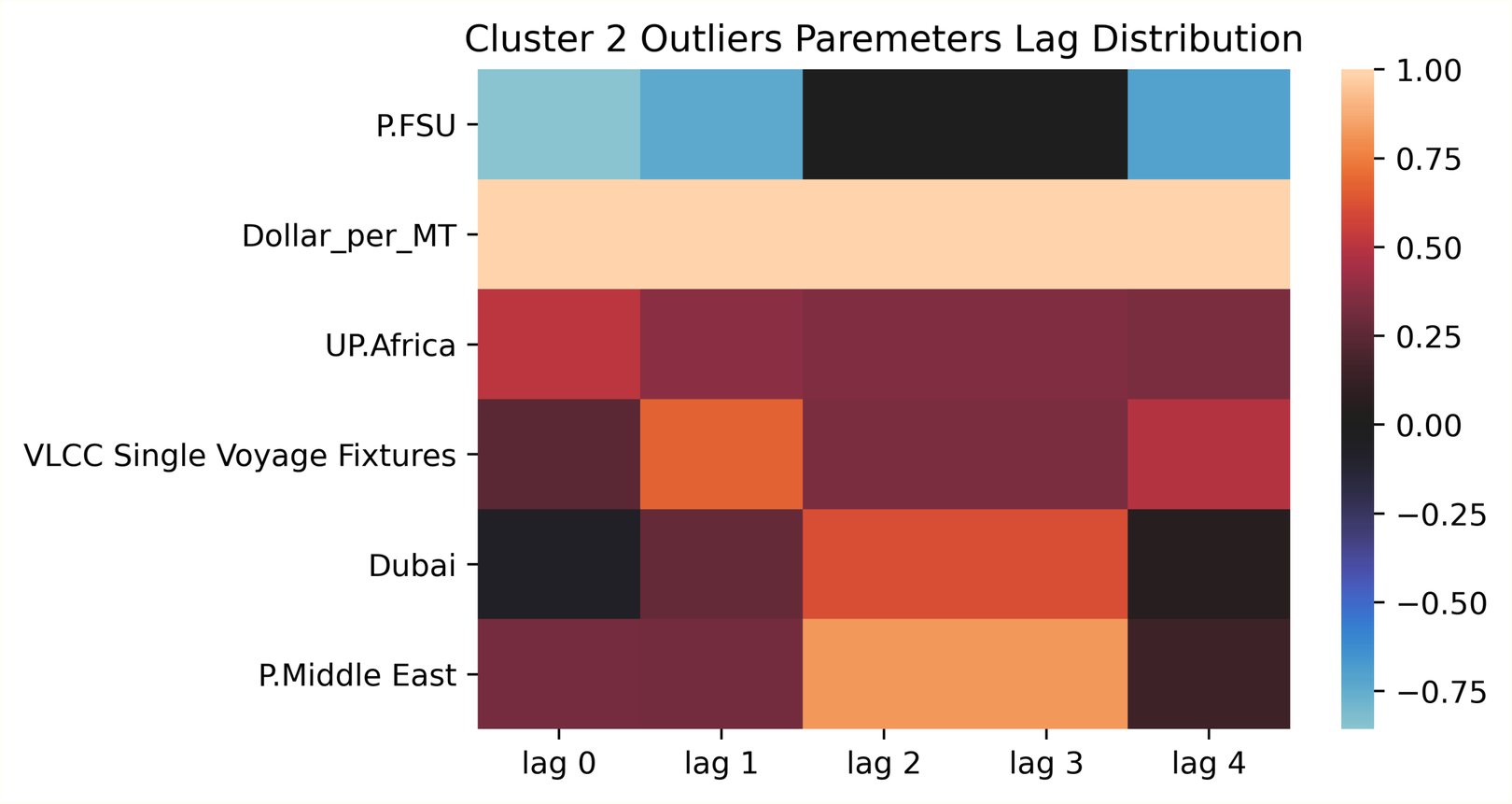
CORRELATED VARIABLES

- We searched for moderately (>0.5) positively or negatively correlated with Dollar per Metric Ton within a time lag of a cluster.
- We then compiled a list of all correlated variables within a particular cluster taken from all time lagged data.
- This process was repeated for every cluster giving us 5 such lists.
- We then tried to find variables which were common among all these lists giving us a final set of variables that were moderately correlated at least on 1 time lag for every cluster.

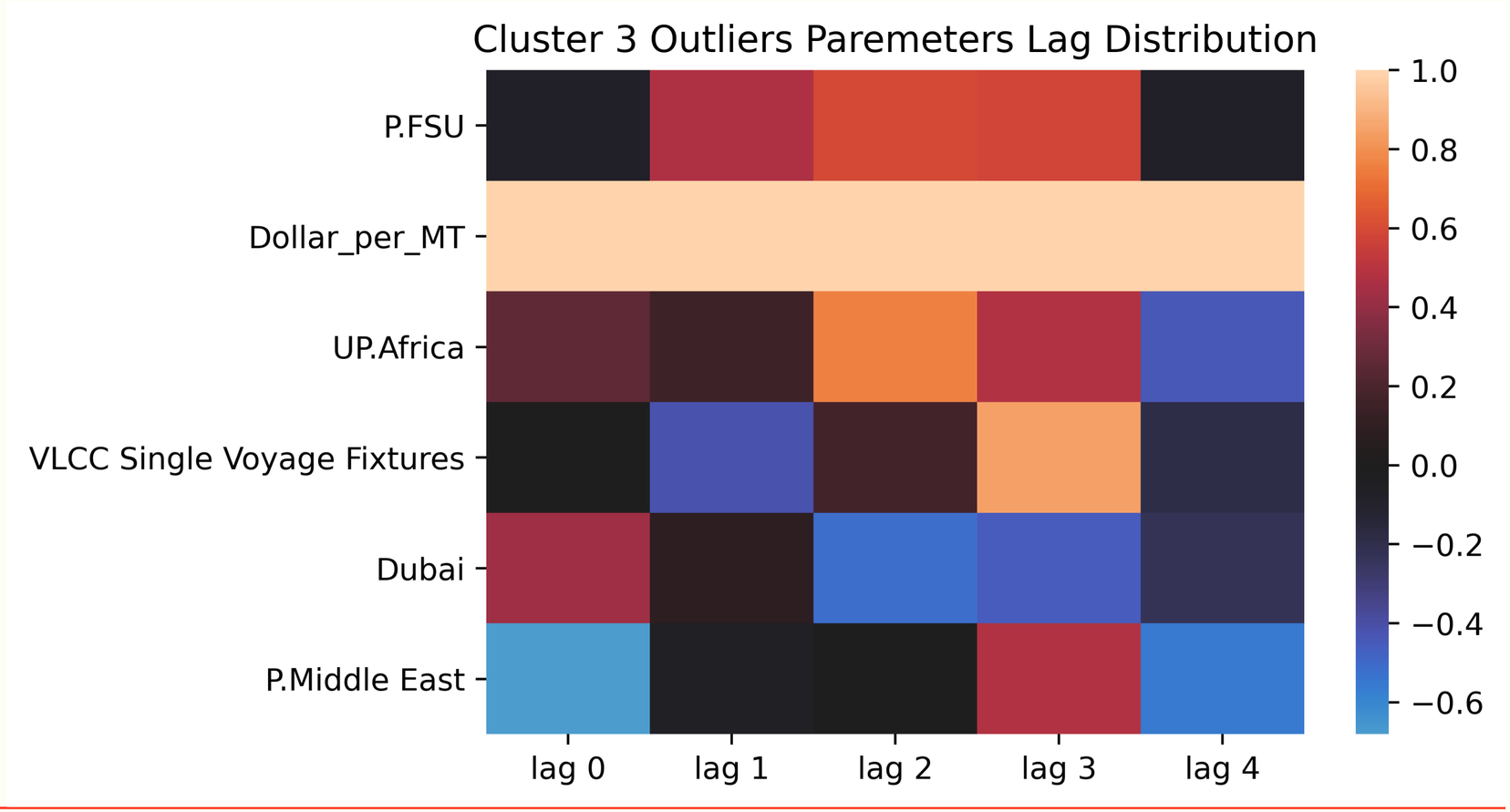
CORRELATED VARIABLES IN CLUSTER 1



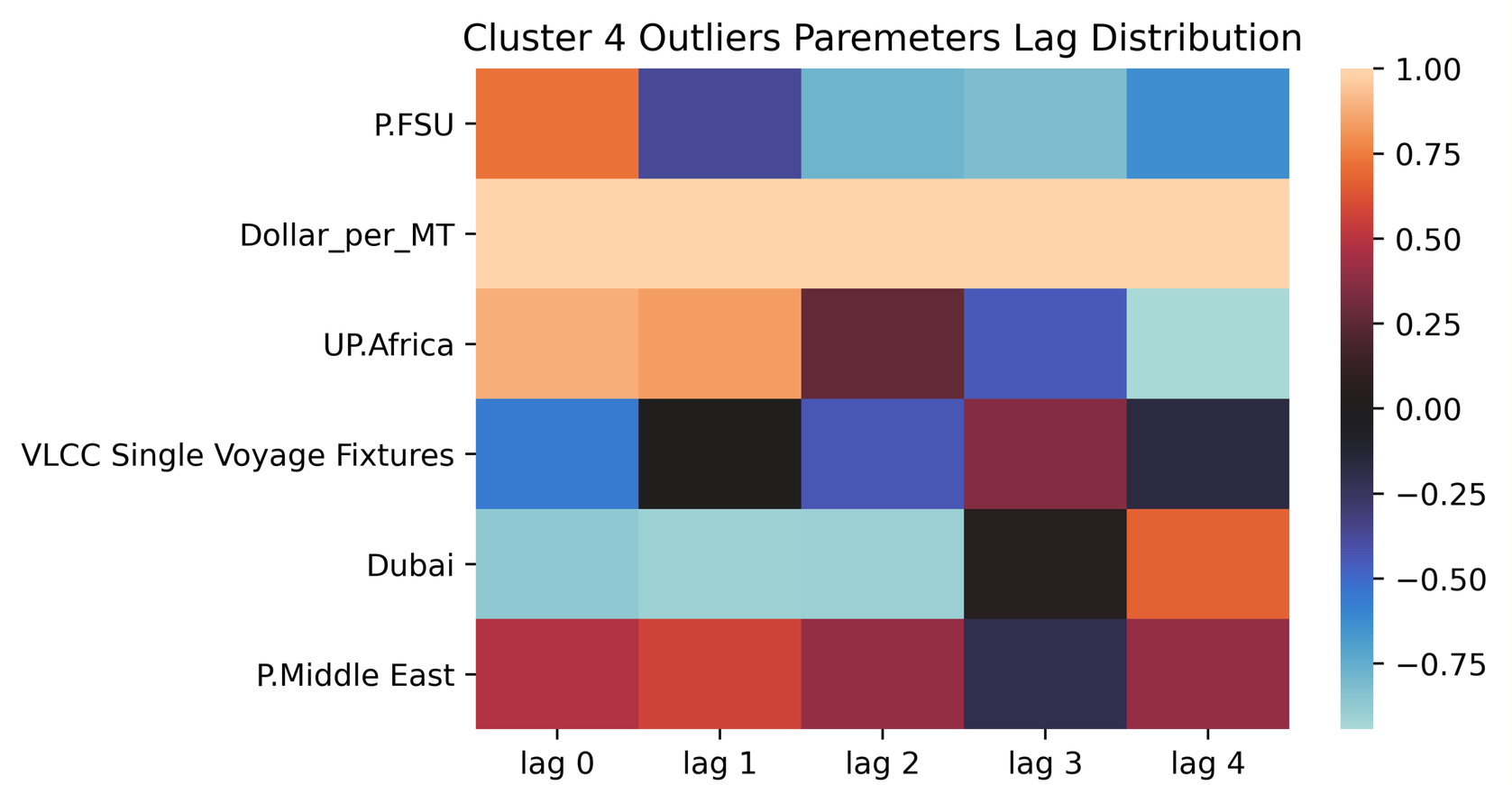
CORRELATED VARIABLES IN CLUSTER 2



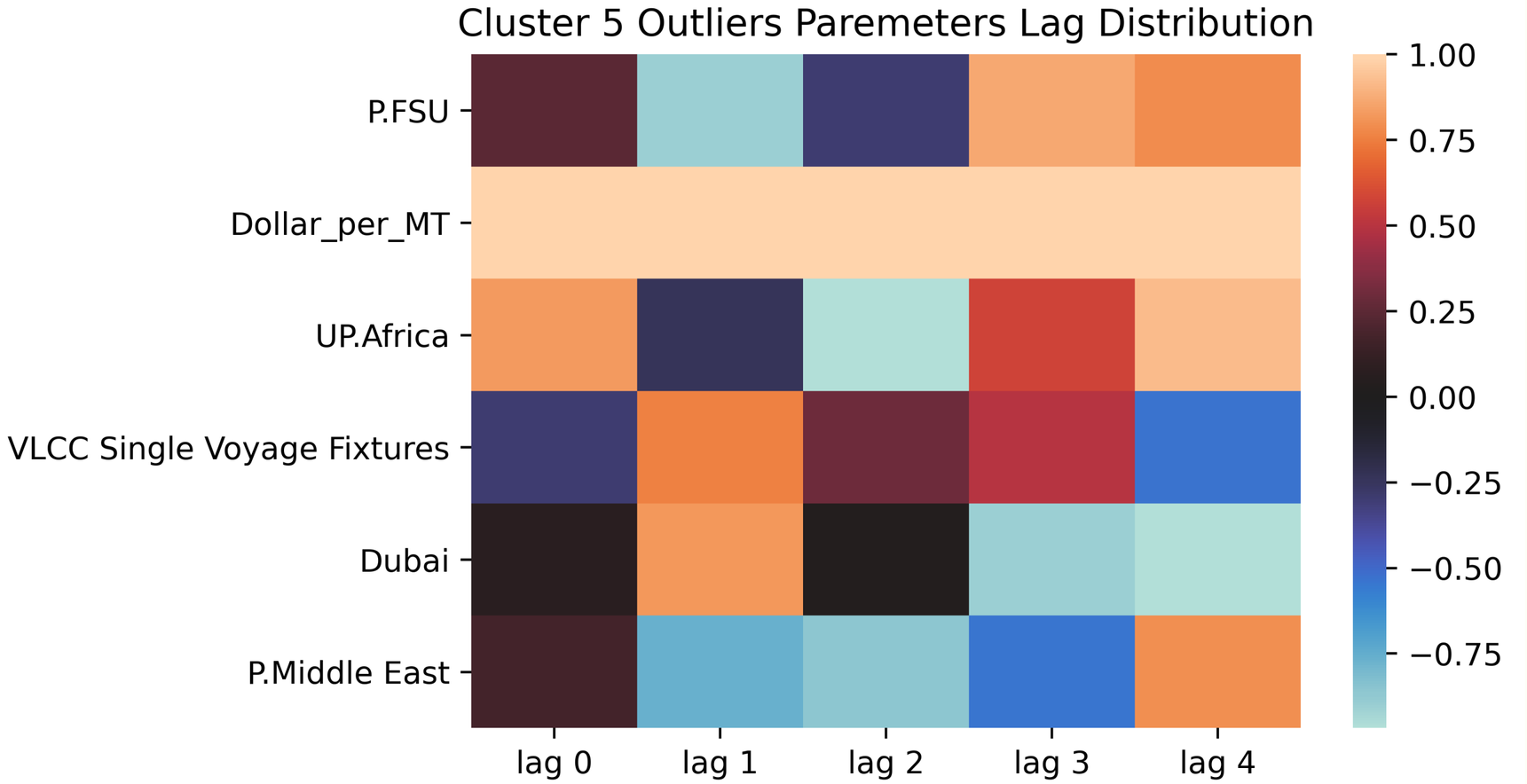
CORRELATED VARIABLES IN CLUSTER 3



CORRELATED VARIABLES IN CLUSTER 4



CORRELATED VARIABLES IN CLUSTER 5



MODELS AND RESULTS

- We used 6 machine learning models on 1 month lagged data to help us find patterns that will allow us to predict future exceptional spot rates values based on present month’s data.
- Our models were tuned using Grid Search based on certain pre decided hyper parameter space with the F1 scoring metrics.
- We used both 4 Fold Stratified Cross Validation and Rolling Window Cross Validation to test every parameter space separately for every model.

Rolling Window Cross Validation

	KNeighborsClassifier	Decision Tree	Logistic Regression	SVM	Random Forest	SDG
F1 Score	0.259259	0.320988	0.345679	0.148148	0.358025	0.407407
Parameters	'metric': 'euclidean', 'n_n	'max_depth': 3,	'C': 1.0, 'penalty': 'l2	'C': 50, 'gai	'max_features': 'sqrt	'learning_

4 Fold Stratisfied Cross Validation

	K neighboors Classfier	Decision Tree	Logistic Regressions	SVM	Random Forest	SDG
F1	0.480769	0.715873	0.287179	0.10101	0.495513	0.702068
Parameters	'metric': 'euclidean', 'n_neighbor	'max_depth': 2, 'min_	'C': 0.01, 'penalty': 'l2	'C': 50, 'gai	'max_features': 'sqrt', 'r	'learning_