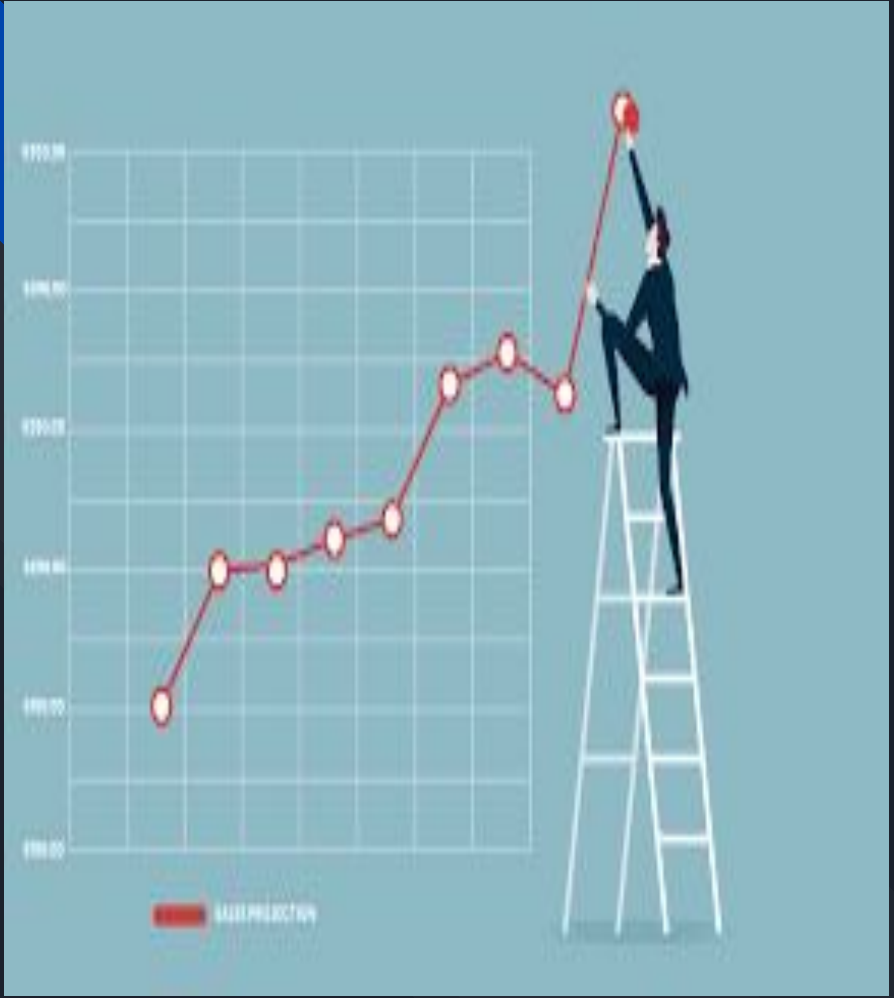


Higher Accuracy Sales Forecasting Through Granular Customer Segmentation

by
Rime Saad

Problem Statement

How would the business achieve deeper insight into customer profiles and higher accuracy sales forecasting

What will be covered in this presentation

- Data cleaning
- Customer cohorts
- Recency, Frequency and Monetary Value and RFM score
- Segmenting customers using kmeans and hierarchical clustering
- Sales Prediction

The data

Shape

- 1,067,371 rows and 8 columns
- 5,835 customers and 4,615 products

Features and data cleansing:

- InvoiceNumber
- Stockcode
- Description
- Quantity: dealing with negative quantity
- Price: dealing with negative prices
- InvoiceDate
- Customer ID: dealing with nan customers
- Country

New feature

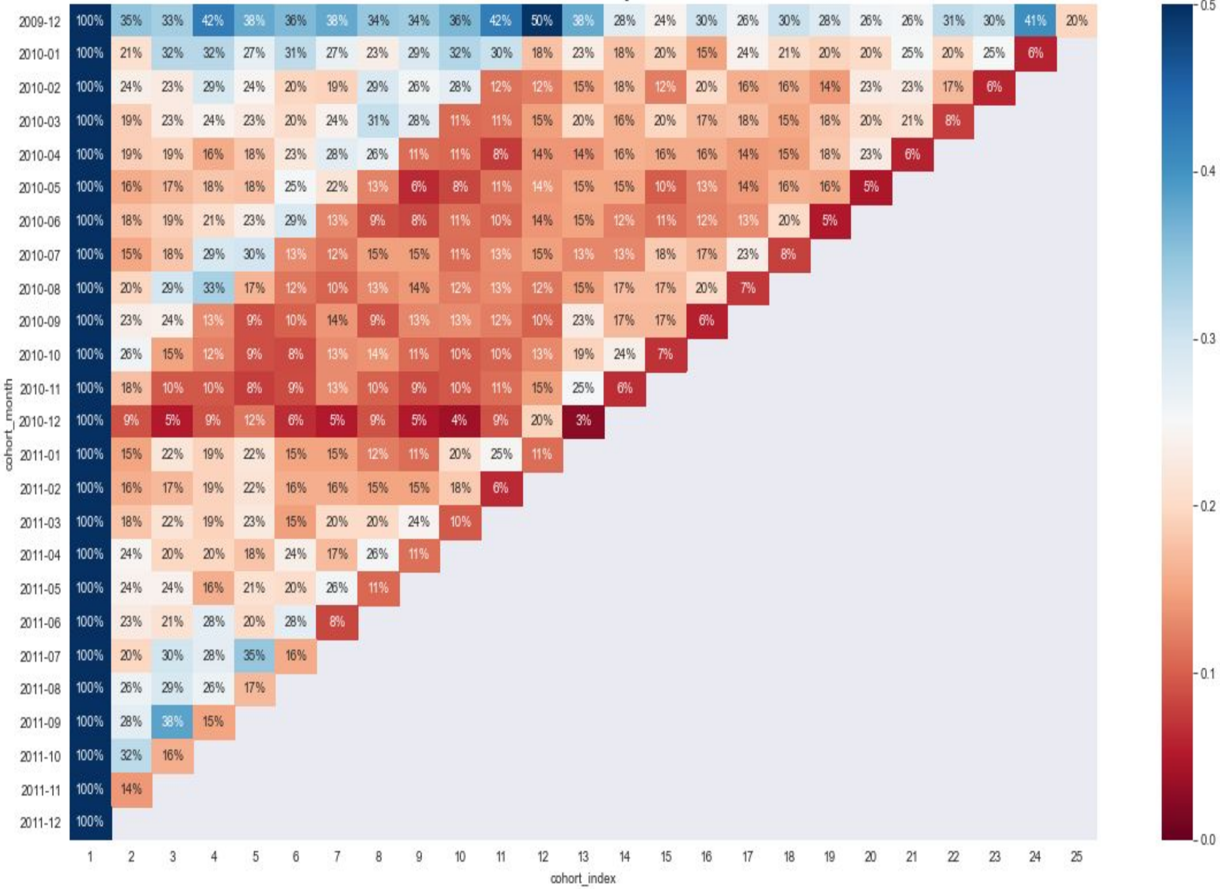
- Sales = quantity*price

Customer Cohorts

Benefits

- Understand how customer behaviors affect the business.
- Reduce customer churn.
- Increase customer lifetime value.
- Increase customer engagement.

Monthly Cohort Analysis



Segmentation Criteria

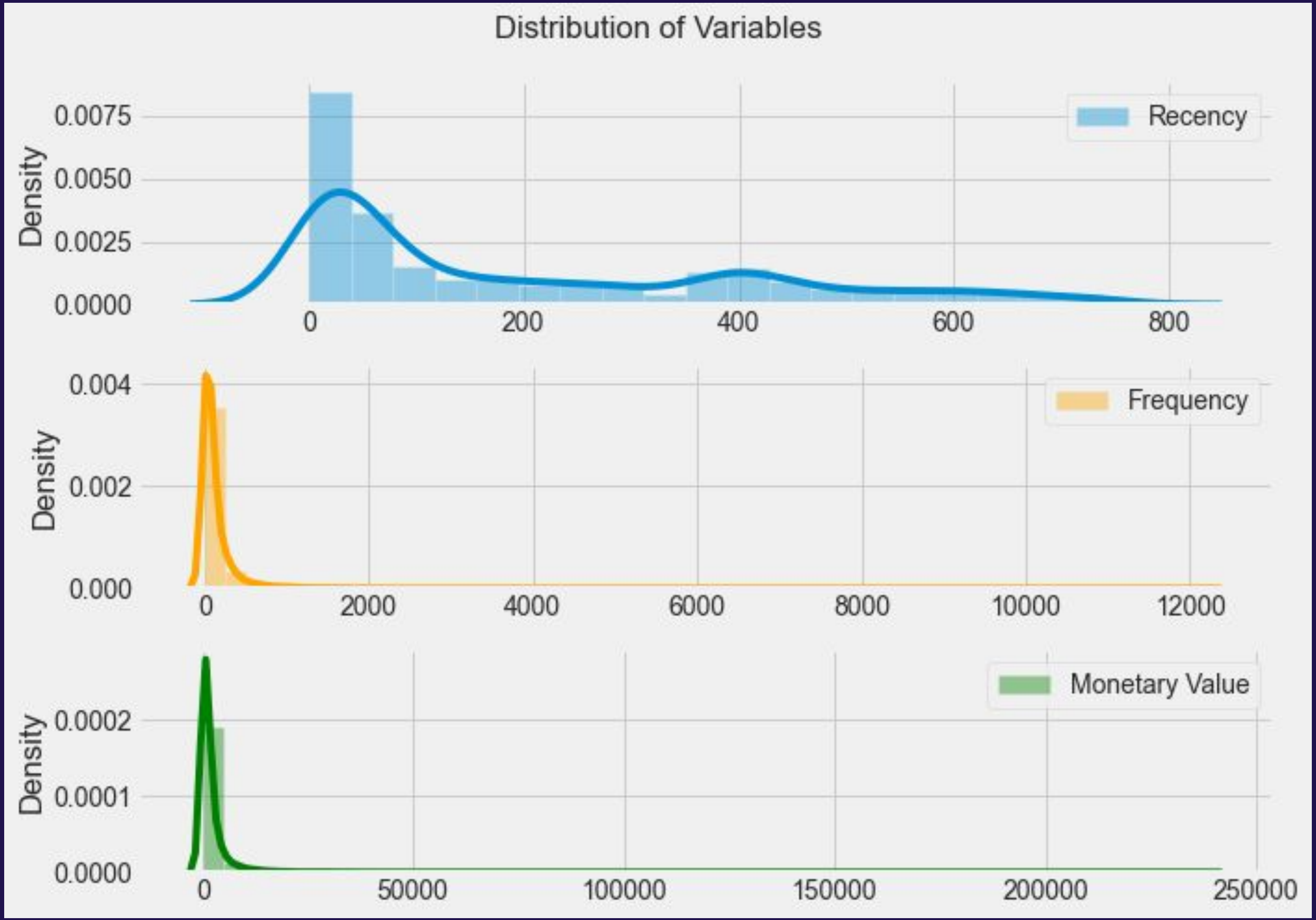
- Recency
- Frequency
- Monetary Value

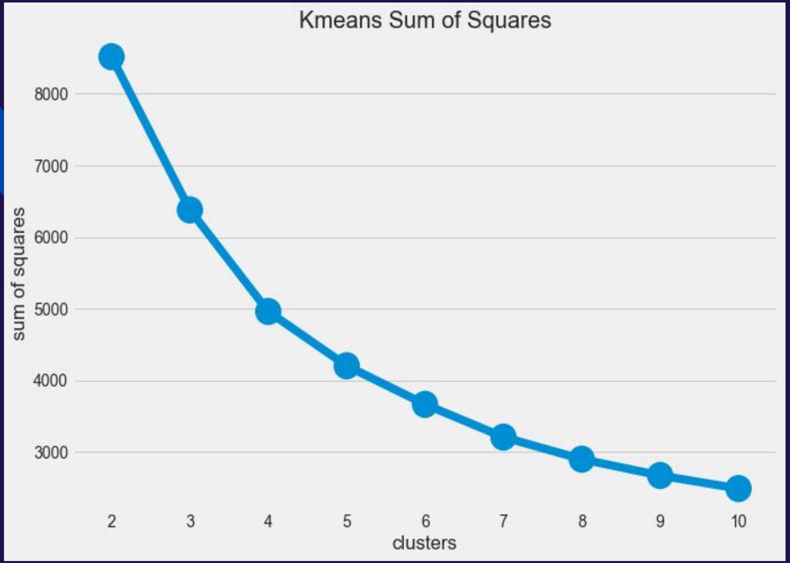
Customer ID	Recency	Frequency	Monetary Value	RFM_segment	RFM_score
12346	529	24	169.36	121	4
12347	2	222	4921.53	444	12
12348	75	46	1658.40	323	8
12349	19	172	3678.69	444	12
12350	310	16	294.40	211	4

Clustering

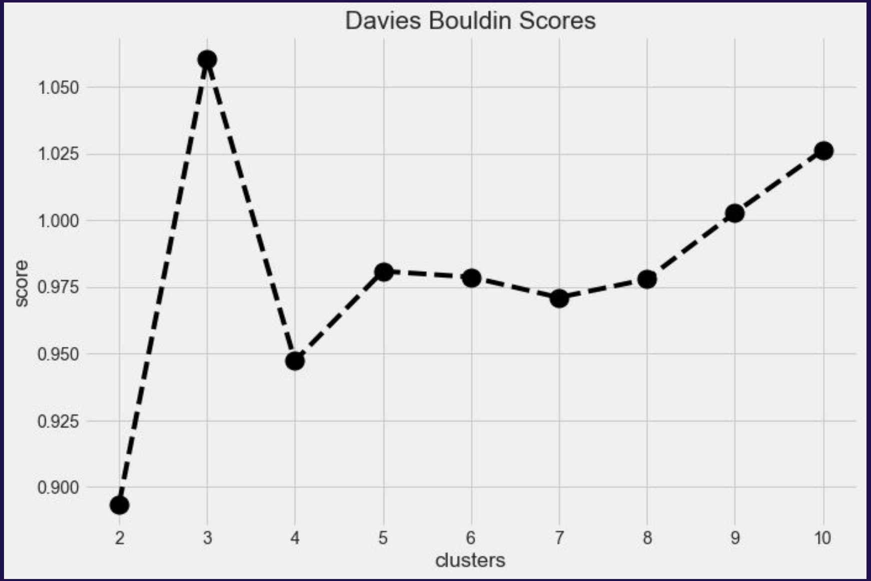
Kmeans

Hierarchical

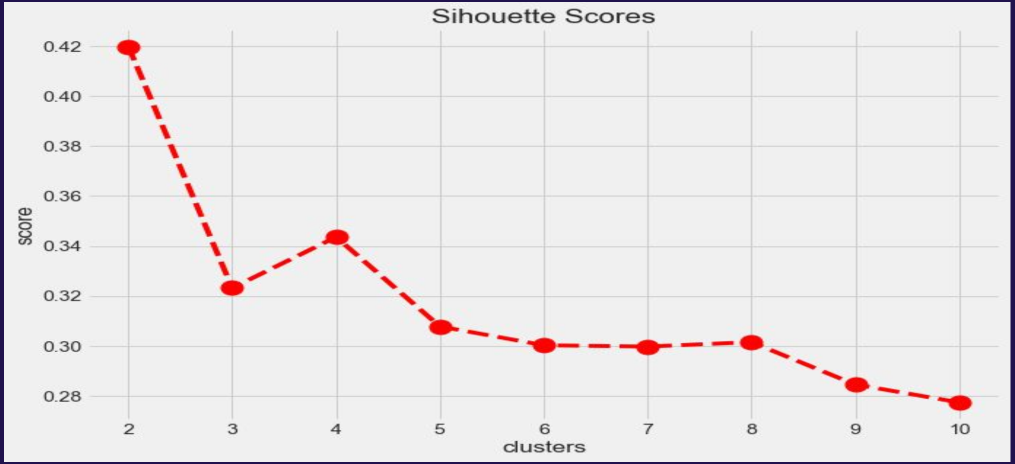

Distribution of Variables
Recency
Density
0.0075
0.0050
0.0025
0.0000
0
200
400
600
800
Frequency
Density
0.004
0.002
0.000
0
2000
4000
6000
8000
10000
12000
Monetary Value
Density
0.0002
0.0001
0.0000
0
50000
100000
150000
200000
250000



Kmeans Sum of Squares

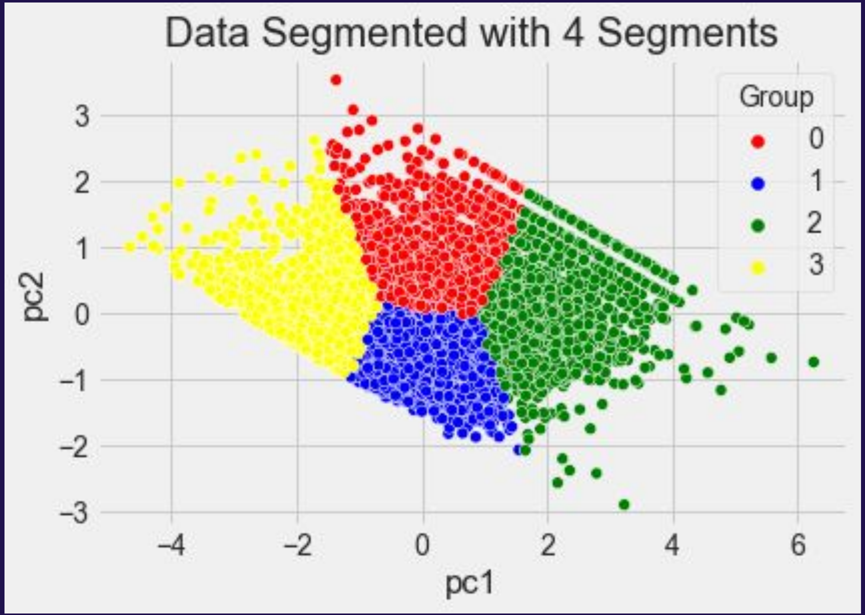


Davies Bouldin Scores

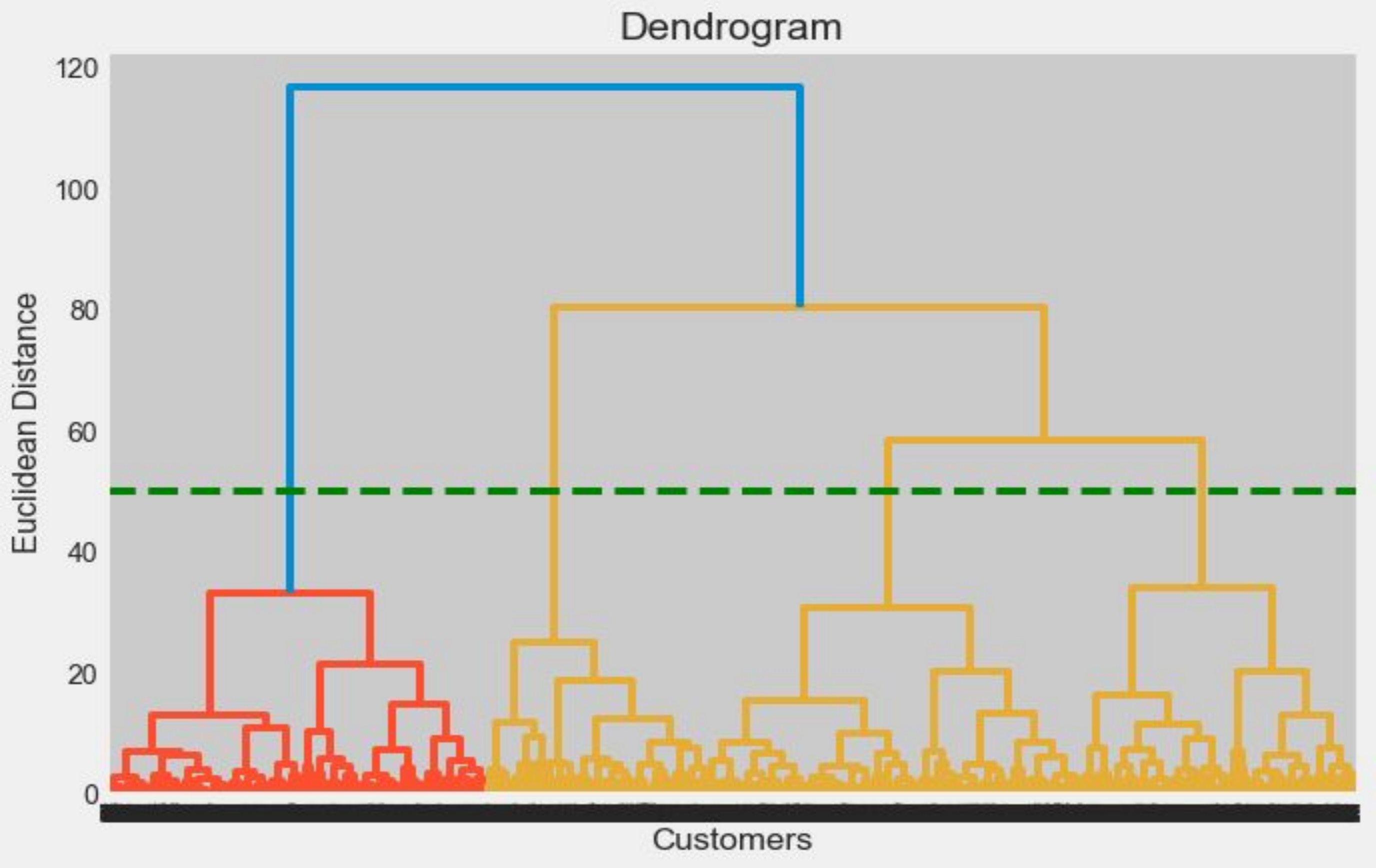


Sihouette Scores

The four segments after PCA



Hierarchical Clustering



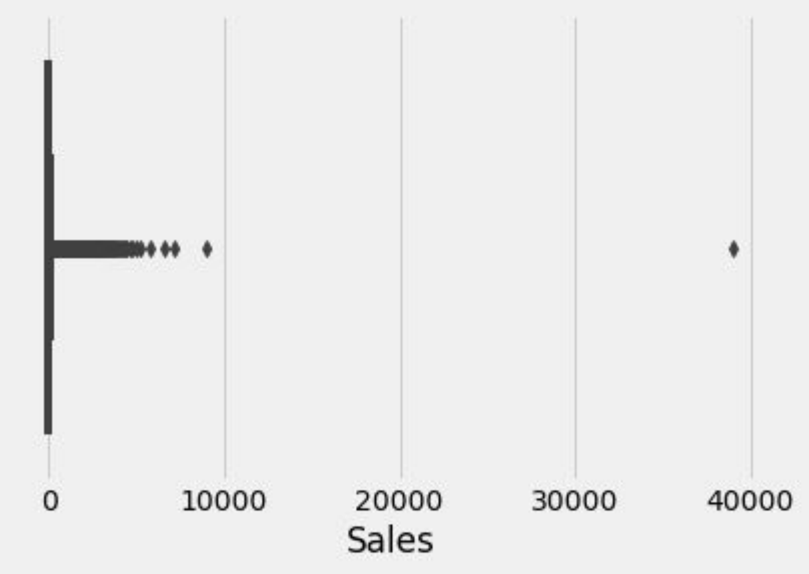
Group	Recency	Frequency	Monetary	Value
	mean	mean	mean	count
1	31.0	55.0	770.0	1207
2	279.0	88.0	1310.0	1568
3	29.0	395.0	6434.0	1311
4	386.01	5.0	235.0	1645

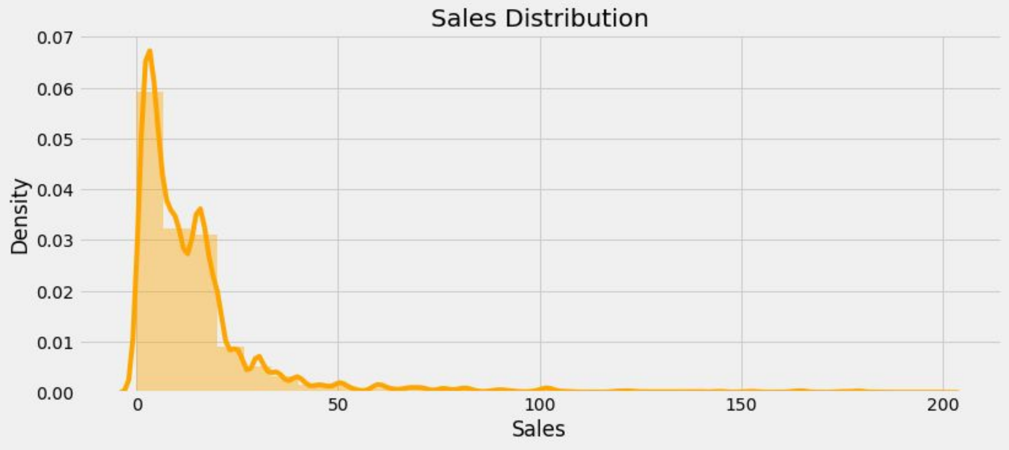
Highest value customers

Sales Forecasting

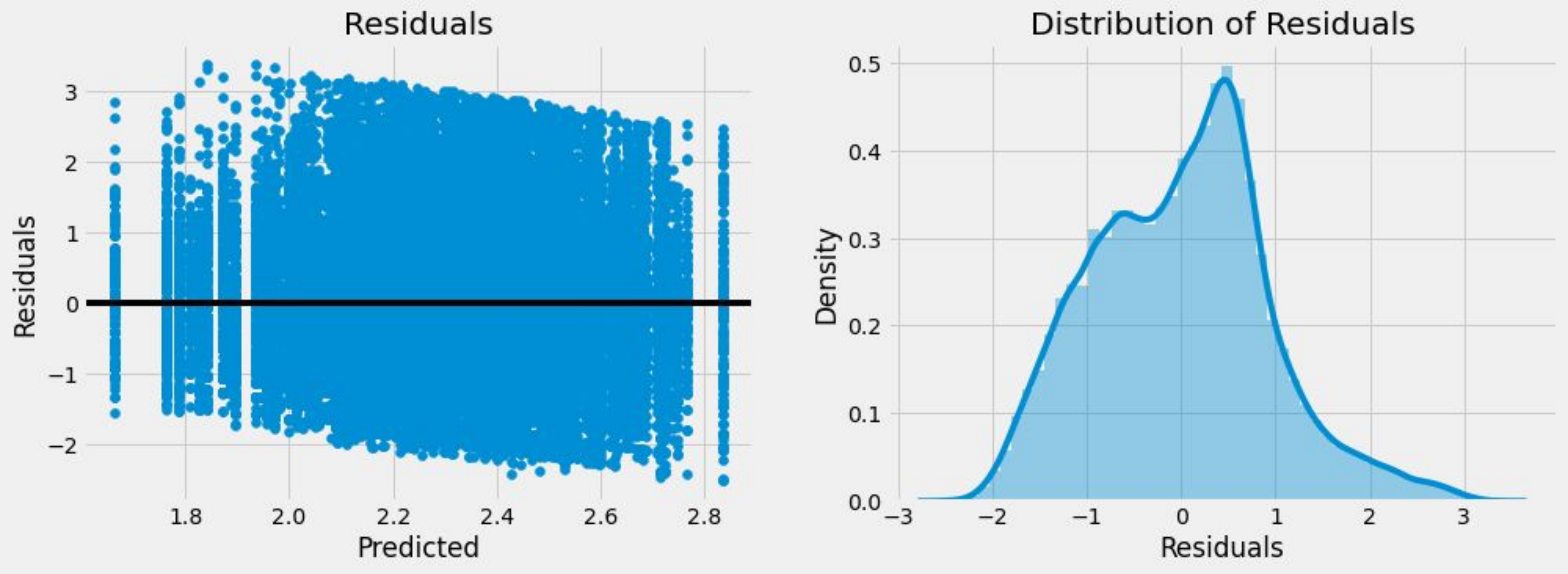
- LinearRegression
- KnearestNeighborRegressor
- RandomForestRegressor
- ARIMA

Sales


0
10000
20000
30000
40000
Sales


Sales Distribution
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
50
100
150
200
Sales

RandomForestRegressor Residuals



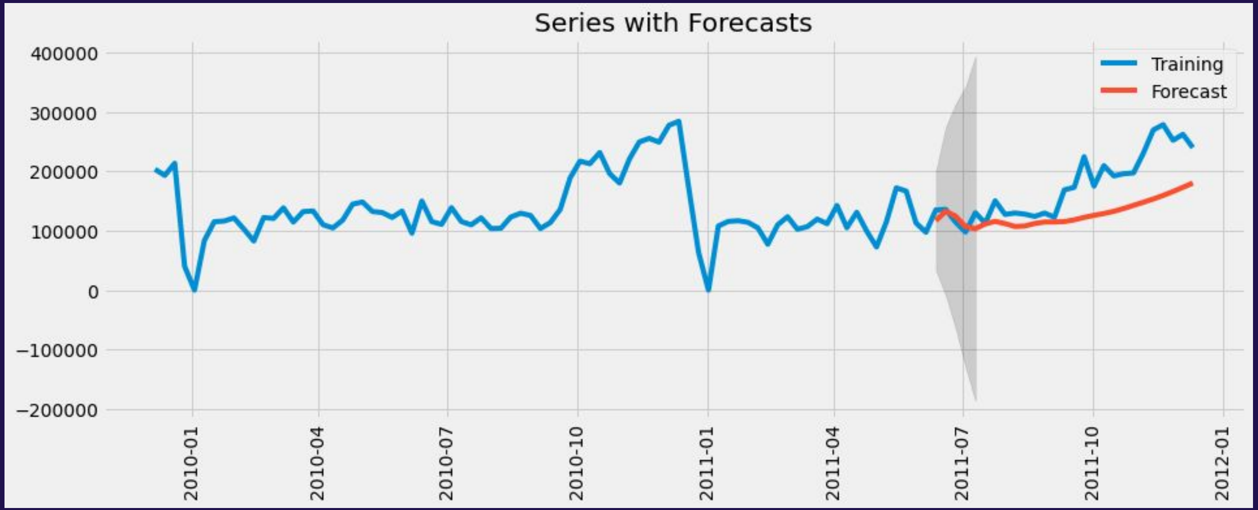
ARIMA

AR == Autoregressive
I == Integrated
MA == Moving Average

Parameters to be tuned

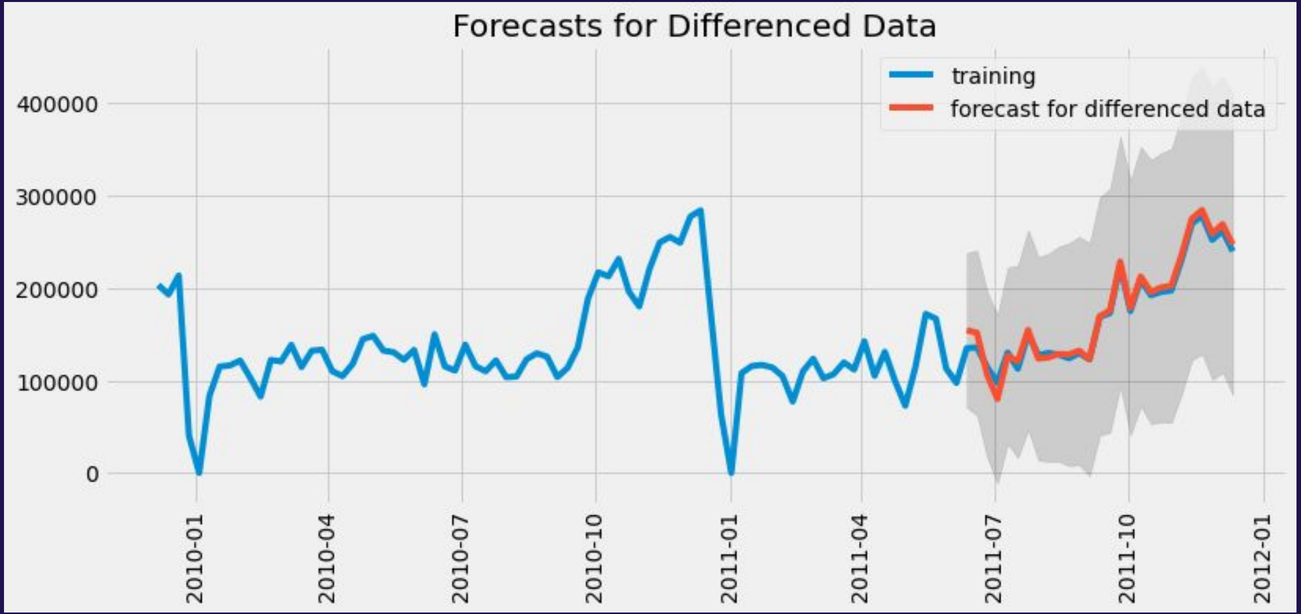
- **p** is the number of lags
- **d** is the minimum number of differencing needed to make the series stationary
- **q** is the number of lagged forecast errors

p=10, d=2, q=0

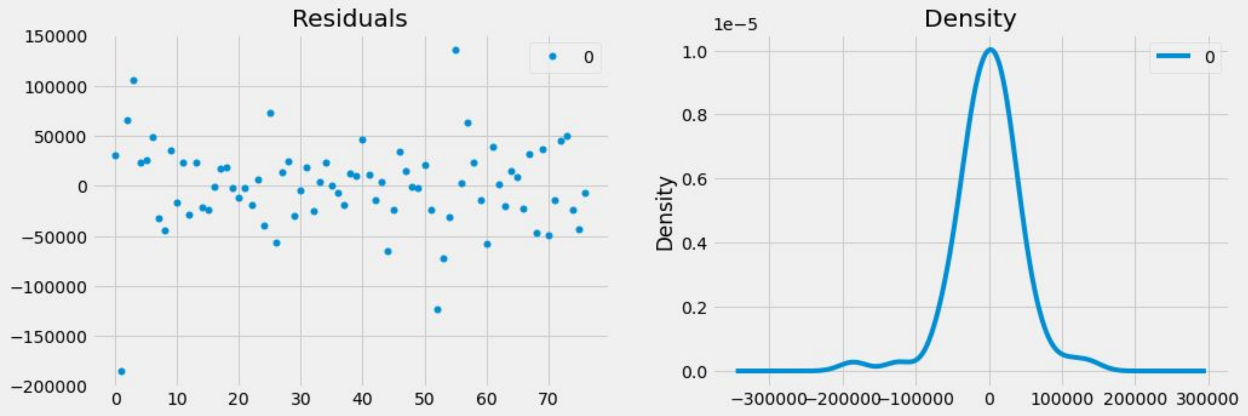


RMSE = 47642

P =4, d=0, q=0



RMSE = 24931


Residuals
150000
100000
50000
0
−50000
−100000
−150000
−200000
0
10
20
30
40
50
60
70
0
Density
1e−5
1.0
0.8
0.6
0.4
0.2
0.0
Density
−300000
−200000
−100000
0
100000
200000
300000
0