Retail Dataset Analysis

Rany Dwi Cahyaningtyas's Portofolio

Rany Dwi Cahyaningtyas

rany.dwi@sci.ui.ac.id

0817 6539 722

I am an undergraduate Statistics student looking for opportunities in **quantitative analysis, machine learning researches**, **Customer Relationship Management**, **actuary,** and **data science** field with a background in **actuarial, data analysis,** and **statistics.**

Table of contents

About the Data

This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.

The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.

Source: https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset

	Definition	Data Type
InvoiceNo	Invoice number	Nominal.
Stock Code	Product (item) code	Nominal
Description	Product (item) name	Nominal
Quantity	The quantities of each product (item) per transaction	Numeric
Invoice Date	Invoice date and time	Numeric
Unit Price	Product price per unit in sterling (Â£).	Numeric
CustomerID	A 5-digit integral number uniquely assigned to each customer	Nominal
Country	The name of the country where a customer resides	Nominal

01

Sales Forecast

Tableau Method for Forecasting

Exponential Smoothing and Trend

- *Exponential Smoothing Models iteratively predict the future value of a series of regular values from the weighted averages of the past values of the series.*

- *Exponential method because the value of each level is affected by each actual value of the previous level that decreases/increases exponentially—newer values are given greater weight*

MAPE	Forecasting power
<10%	Highly accurate forecasting
10%~20%	Good forecasting
20%~50%	Reasonable forecasting
>50%	Weak and inaccurate forecasting

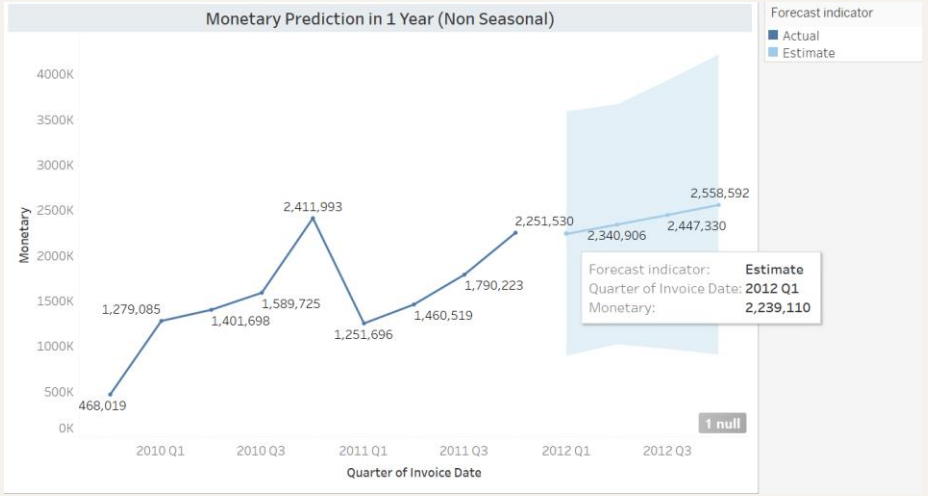
Source: Lewis (1982)

Forecast Revenue in 4 Quartal (Model 1)

Using ARIMA (Trend Multiplicative)

Estimate Revenue in:

Q1 2012	—	2.239.110
Q2 2012	—	2.340.906
Q3 2012	—	2.447.330
Q4 2012	—	2.558.592



9.585.938

Total Revenue in 2012

Analysis Model 1

All forecasts were computed using exponential smoothing.

Sum of Monetary

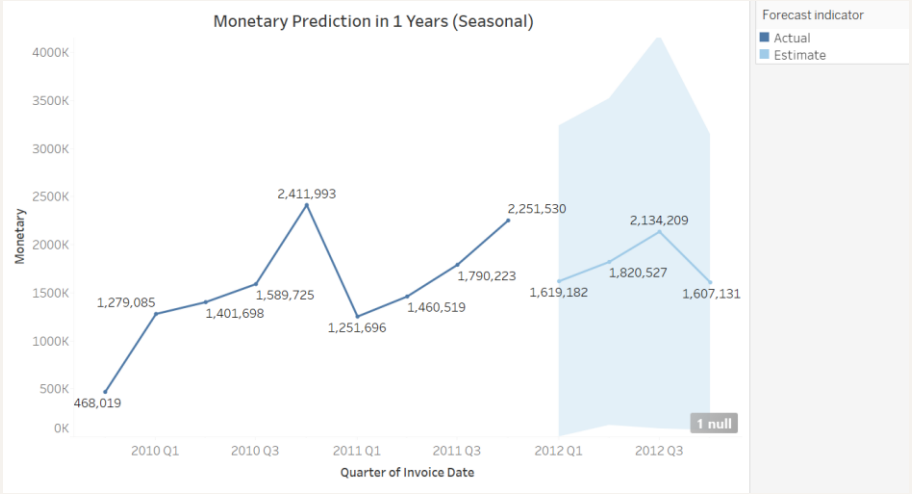
Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Multiplicative	Multiplicative	None	548,474	416,289	0.81	42,0%	248	0.211	0.500	0.000

Forecast Revenue in 4 Quartal (Model 2)

Using SARIMA (Seasonal Multiplicative)

Estimate Revenue in:

Q1 2012	—	1.619.182
Q2 2012	—	1.820.527
Q3 2012	—	2.134.209
Q4 2012	—	1.607.131



7.181.049

Total Revenue in 2012

Analysis Model 2

All forecasts were computed using exponential smoothing.

Sum of Monetary

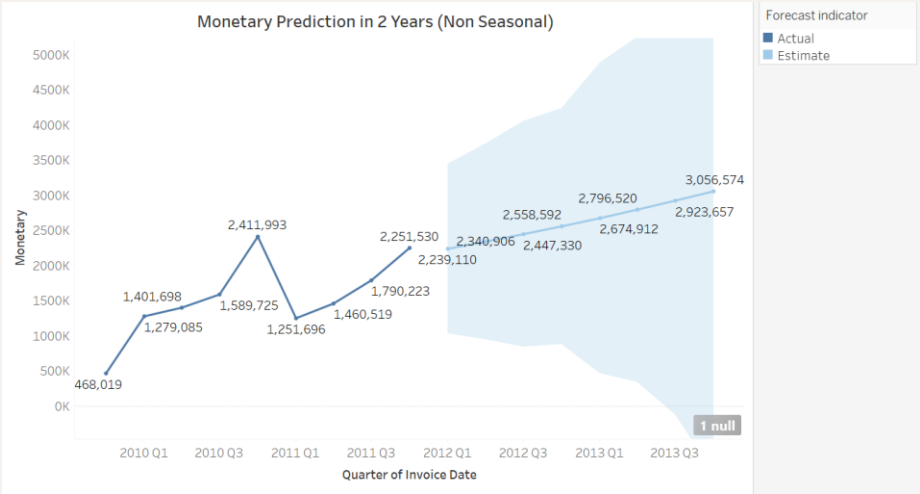
Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Multiplicative	None	Multiplicative	593,008	442,668	0.93	34,2%	253	0.274	0.000	0.000

Forecast Revenue in 8 Quartal (Model 3)

Using ARIMA (Trend Multiplicative)

Estimate Revenue in:

Q1 2012	—	2.239.110
Q2 2012	—	2.340.906
Q3 2012	—	2.447.330
Q4 2012	—	2.558.592
Q1 2013	—	2.674.912
Q2 2013	—	2.796.520
Q3 2013	—	2.923.657
Q4 2013	—	3.056.574



9.585.938
Total Revenue in 2012

11.451.663
Total Revenue in 2013

Analysis Model 3

All forecasts were computed using exponential smoothing.

Sum of Monetary

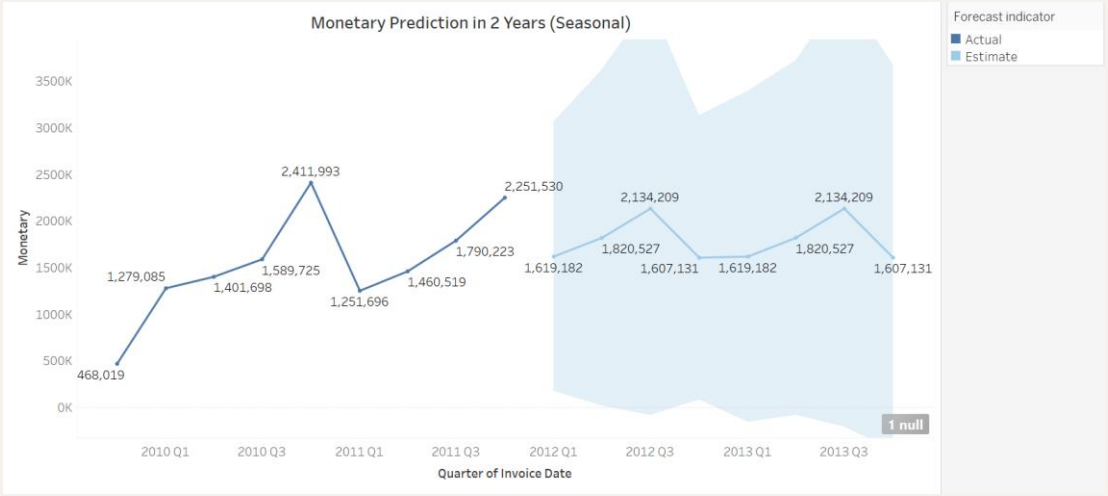
Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Multiplicative	Multiplicative	None	548,474	416,289	0.81	42,0%	248	0.211	0.500	0.000

Forecast Revenue in 8 Quartal (Model 4)

Using SARIMA (Seasonal Multiplicative)

Estimate Revenue in:

Q1 2012	—	1.619.182
Q2 2012	—	1.820.527
Q3 2012	—	2.134.209
Q4 2012	—	1.607.131
Q1 2013	—	1.619.182
Q2 2013	—	1.820.527
Q3 2013	—	2.134.209
Q4 2013	—	1.607.131



1.607.131
Total Revenue in 2012

1.607.131
Total Revenue in 2013

Analysis Model 4

All forecasts were computed using exponential smoothing.

Sum of Monetary

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Multiplicative	None	Multiplicative	593,008	442,668	0.93	34,2%	253	0.274	0.000	0.000

Model Selection for Forecast

	MAPE
Model 1 (Forecast Revenue in 4 Quartal – Non Seasonal)	42%
Model 2 (Forecast Revenue in 4 Quartal – Seasonal)	34,2%
Model 3 (Forecast Revenue in 8 Quartal – Non Seasonal)	42%
Model 4 (Forecast Revenue in 8 Quartal – Seasonal)	34,2%

We choose Model 2 dan Model 4 as a Correct Model

02

Regression Model

OLS Regression

2010 Monetary, Recency, and avg_order_cost are **significant variable** for predict **Monetary in 2011**

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.853
Model:	OLS	Adj. R-squared (uncentered):	0.853
Method:	Least Squares	F-statistic:	2602.
Date:	Wed, 23 Mar 2022	Prob (F-statistic):	0.00
Time:	14:12:53	Log-Likelihood:	-16805.
No. Observations:	1797	AIC:	3.362e+04
Df Residuals:	1793	BIC:	3.364e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
2010 Monetary	0.8821	0.012	75.010	0.000	0.859	0.905
Recency	2.2981	0.691	3.326	0.001	0.943	3.653
Frequency	-17.0173	12.046	-1.413	0.158	-40.643	6.608
avg_order_cost	-0.2666	0.140	-1.908	0.057	-0.541	0.007

Omnibus:	3103.582	Durbin-Watson:	2.024
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8616065.462
Skew:	11.095	Prob(JB):	0.00
Kurtosis:	341.497	Cond. No.	1.42e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.42e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Best Model from OLS Regression with p-value < $\alpha = 0.90$

$$\hat{y}_i = \beta_0 + \beta_1 x_{2010\ Monetary} + \beta_2 x_{recency} + \beta_3 x_{avg\_order\_cost}$$

$$\hat{y}_i = 0.8821 x_{2010\ Monetary} + 2.2981 x_{recency} \pm 0.266 x_{avg\_order\_cost}$$

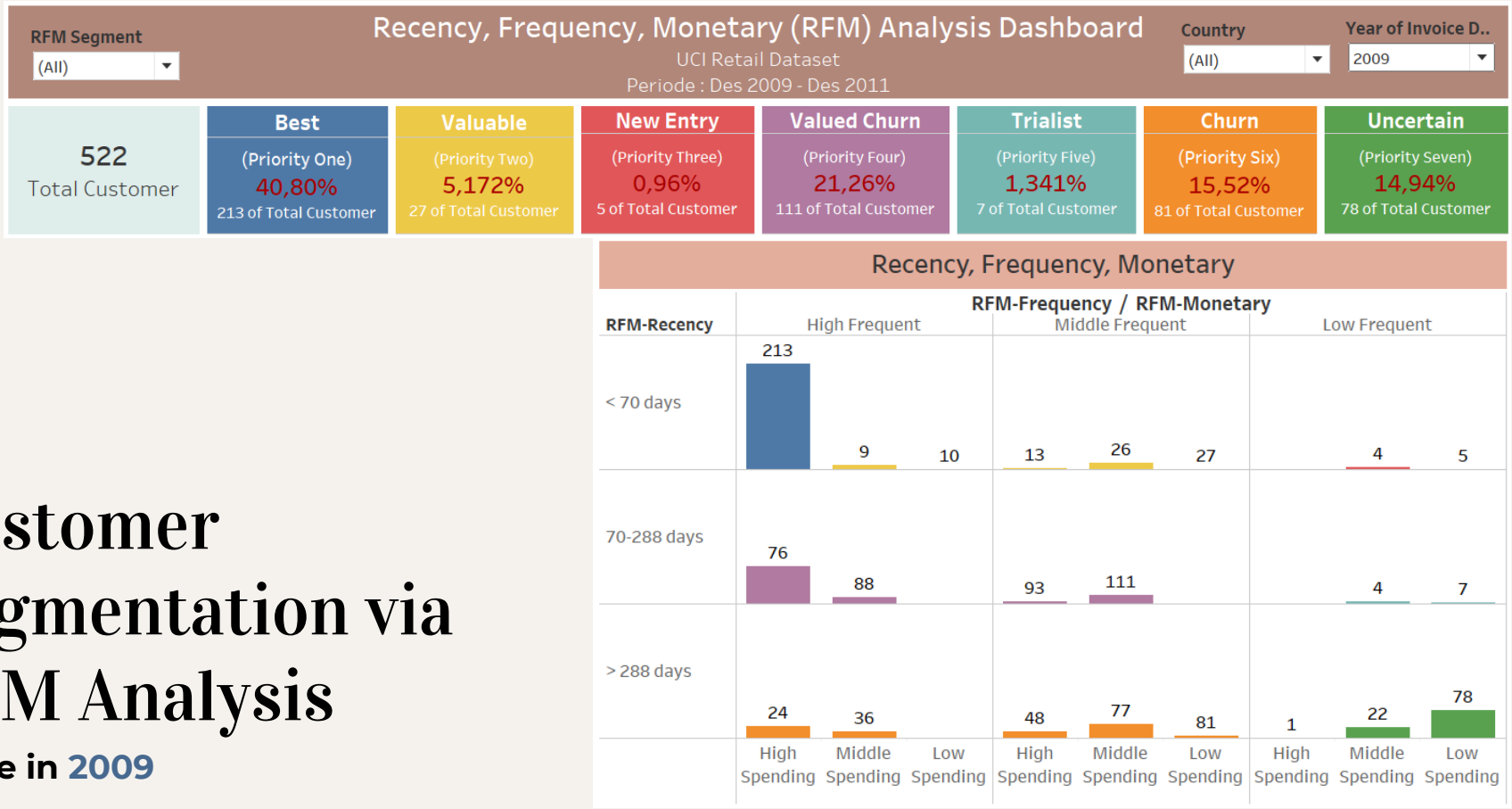
- Where the F-statistic model = $0.00 < \alpha = 0.90$ so the model is quite useful for predicting monetary in 2011
- The value of $R^2_{adj} = 0.853$ implies that 85,3% variation of the data can be explained by the model

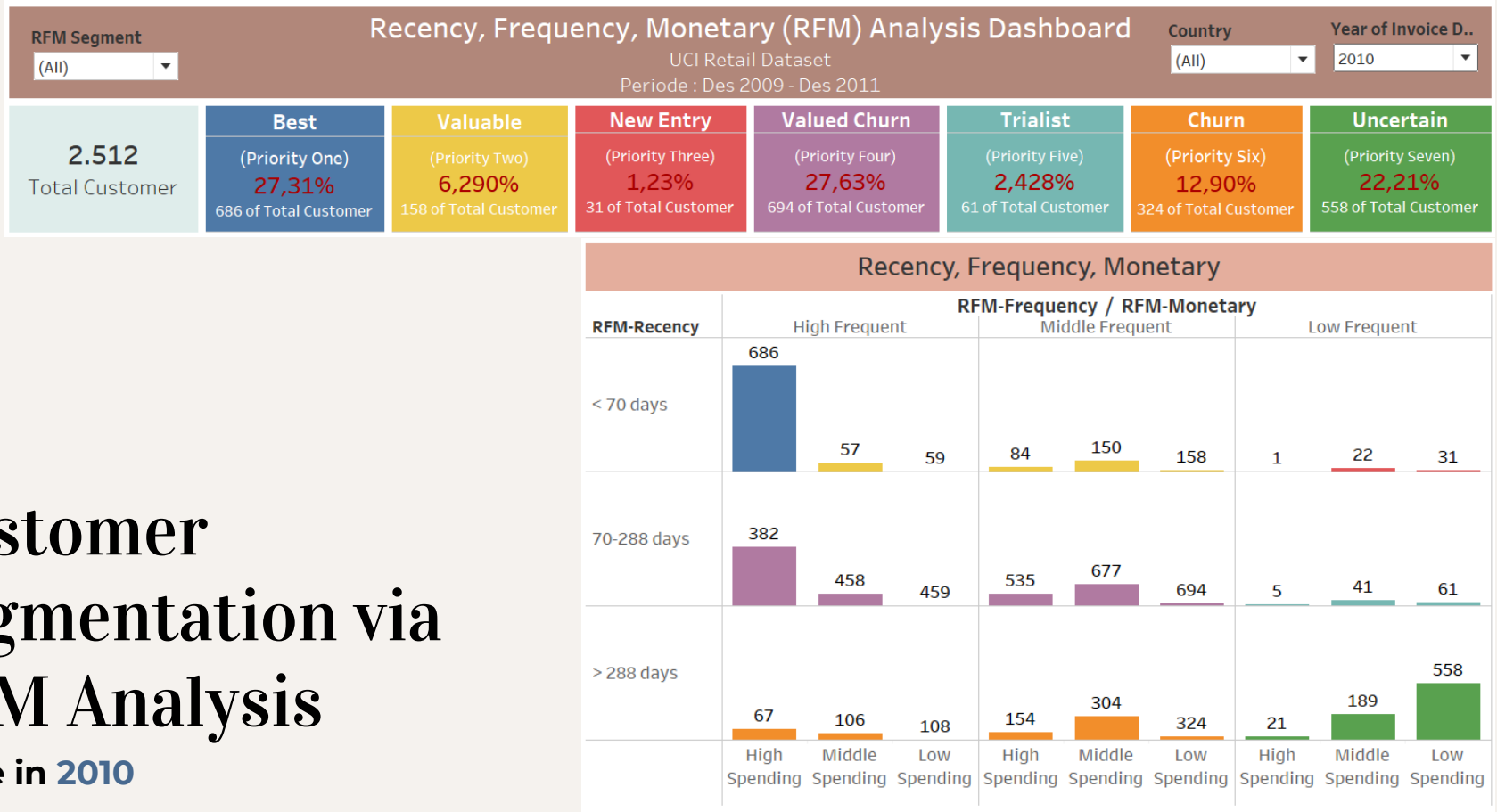
03

RFM Analysis

Customer Segmentation via RFM Analysis

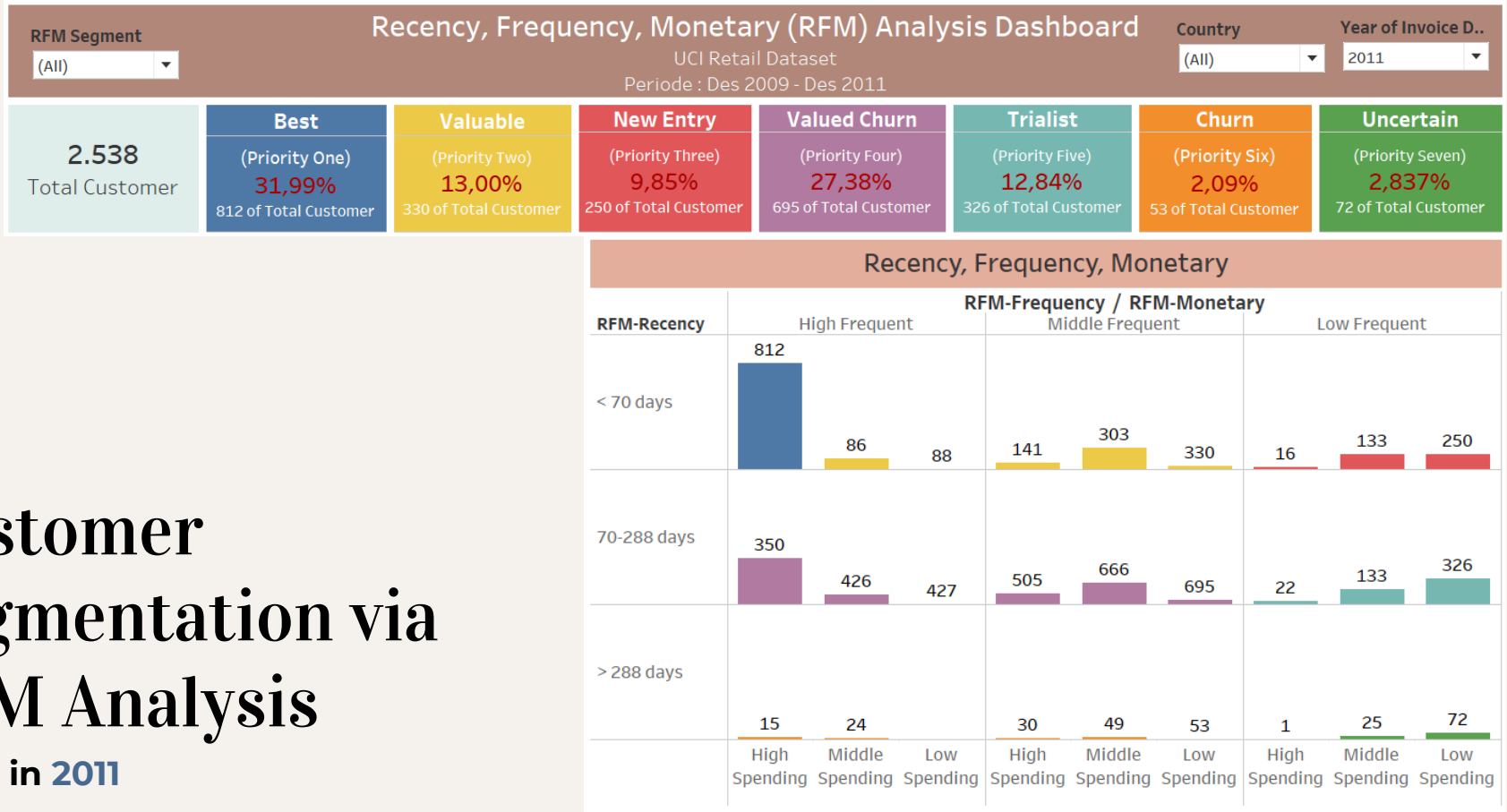
Case in 2009





Customer Segmentation via RFM Analysis

Case in 2010



Customer Segmentation via RFM Analysis

Case in 2011

Detail Customer ID in each Segment

1

Detail Customer Segment **Best**

RFM Segment	Customer ID	Frequency	Monetary	Recency
Best	12347	8	5,476	23
	12352	6	2,034	57
	12358	5	3,887	22
	12362	11	4,943	24
	12388	7	2,760	36
	12395	11	3,898	40
	12408	7	4,341	53
	12415	17	102,874	45
	12417	15	4,790	24
	12423	8	2,096	21

2

Detail Customer Segment **Valuable**

RFM Segment	Customer ID	Frequency	Monetary	Recency
Valuable	12381	5	629	25
	12427	3	826	42
	12428	4	6,541	46
	12444	4	4,248	42
	12464	4	867	31
	12504	4	779	39
	12517	4	2,250	49
	12541	3	981	50
	12597	4	3,727	40
	12610	4	1,530	43

3

Detail Customer Segment **New Entry**

RFM Segment	Customer ID	Frequency	Monetary	Recency
New Entry	12357	2	18,288	54
	12367	1	169	25
	12384	2	585	49
	12397	2	1,365	56
	12442	1	172	24
	12462	2	1,190	39
	12478	1	681	24
	12491	1	460	59
	12508	2	398	47
	12526	2	797	21

4

Detail Customer Segment **Valued Churn**

RFM Segment	Customer ID	Frequency	Monetary	Recency
Valued Churn	12359	6	6,711	78
	12380	6	5,892	77
	12393	6	2,347	93
	12399	6	1,497	140
	12407	5	1,708	70
	12409	6	19,008	99
	12413	4	999	87
	12414	3	682	238
	12422	11	4,415	116
	12450	3	434	177

Detail Customer ID in each Segment

5

Detail Customer Segment **Trialist**

RFM Segment	Customer ID	Frequency	Monetary	Recency
Trialist	12348	2	1,260	269
	12363	2	552	130
	12371	2	2,457	80
	12375	1	190	119
	12378	2	5,374	150
	12394	2	1,272	84
	12420	1	600	84
	12461	2	275	115
	12519	2	727	84
	12534	1	1,050	151

6

Detail Customer Segment **Churn**

RFM Segment	Customer ID	Frequency	Monetary	Recency
Churn	12623	5	2,533	297
	12651	3	246	355
	12755	9	5,227	301
	12866	3	1,212	304
	12933	5	1,172	304
	12956	4	488	327
	13093	36	43,148	296
	13382	5	1,317	301
	13497	4	867	303
	13584	3	392	327

7

Detail Customer Segment **Uncertain**

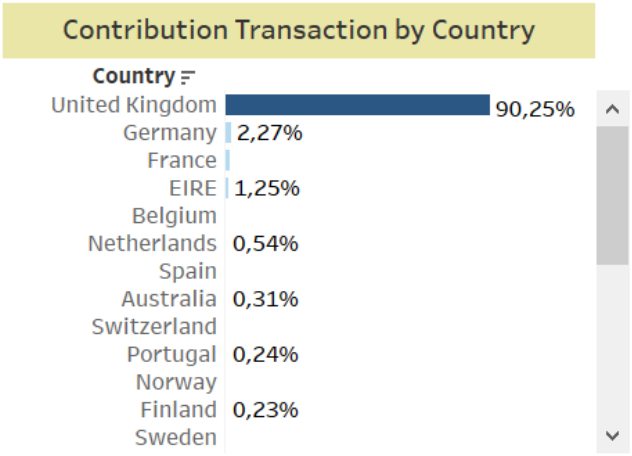
RFM Segment	Customer ID	Frequency	Monetary	Recency
Uncertain	12350	1	239	331
	12373	2	803	332
	12386	2	402	358
	12401	1	84	324
	12410	2	1,014	329
	12489	1	335	357
	12501	1	2,169	357
	12559	2	562	331
	12561	1	239	323
	12735	2	780	336

04

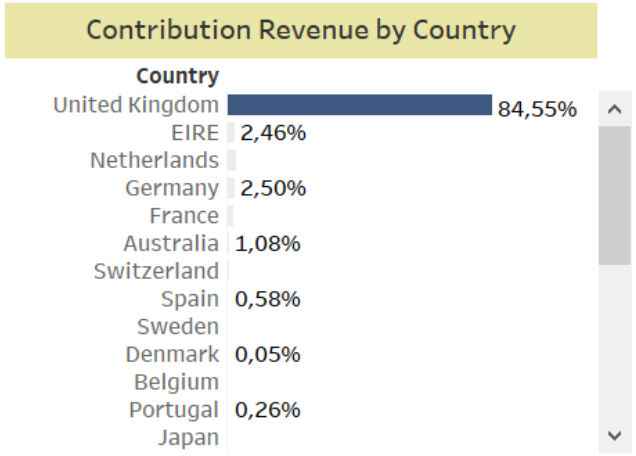
Retail Data Report

Contribution by Country

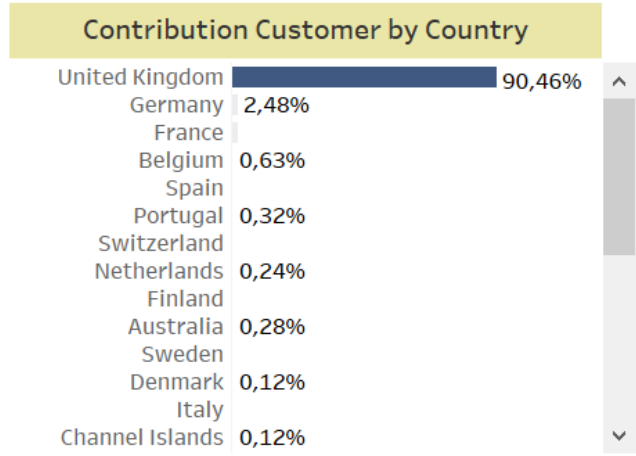
Transaction, Customer, and Revenue in 2011



About **90,25%** total transactions come from **United Kingdom**. Next up is **Germany** with **2,27%** of total transactions



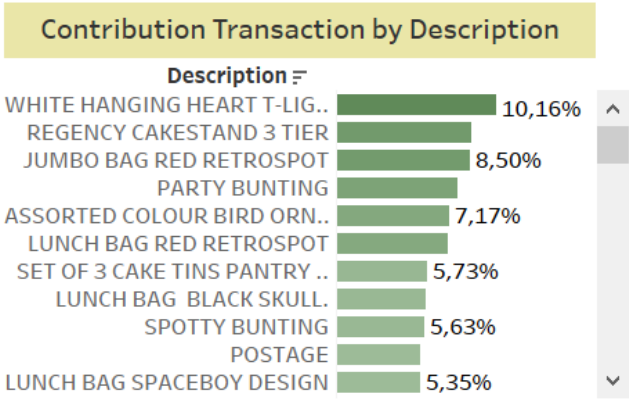
About **84,55%** total revenue come from **United Kingdom**. Next up is **Ireland** with **2,46%** of total revenue

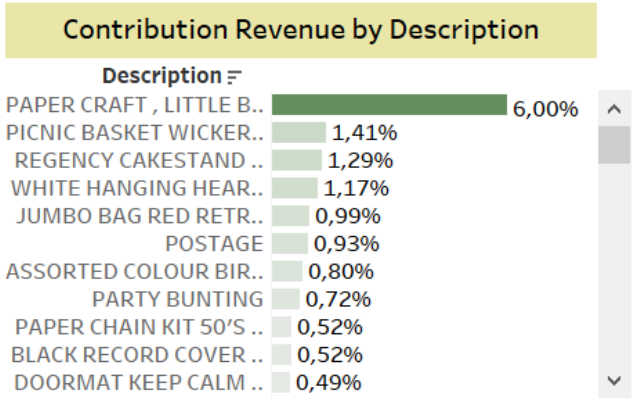


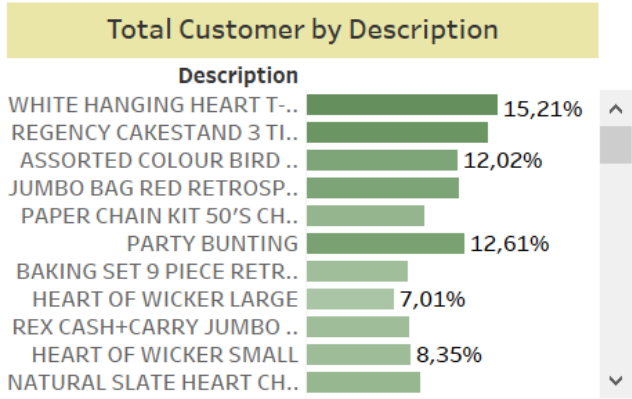
About **90,46%** total customer come from **United Kingdom**. Next up is **Germany** with **2,48%** of total customer

Contribution by Description

Total Transaction, Customer, and Revenue in 2011







White hanging heart T-light Holder 587 accounted for **10,16%** of Total Transaction in Description

Paper Craft, Little Birdy accounted for **6%** of Total Revenue in Description

White hanging heart T-light Holder 587 accounted for **15,21%** of Total Customer in Description