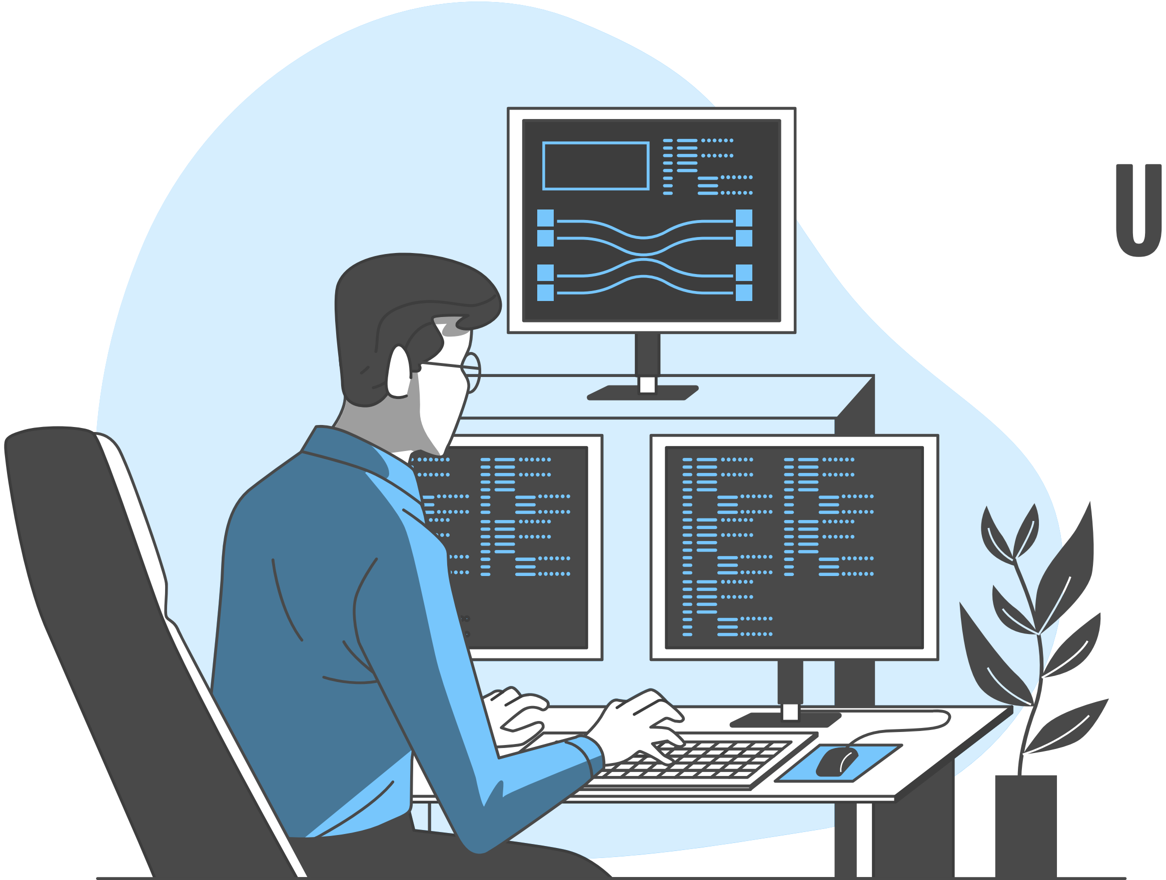


UK E-Commerce Data Analysis

- Giang Nguyễn -

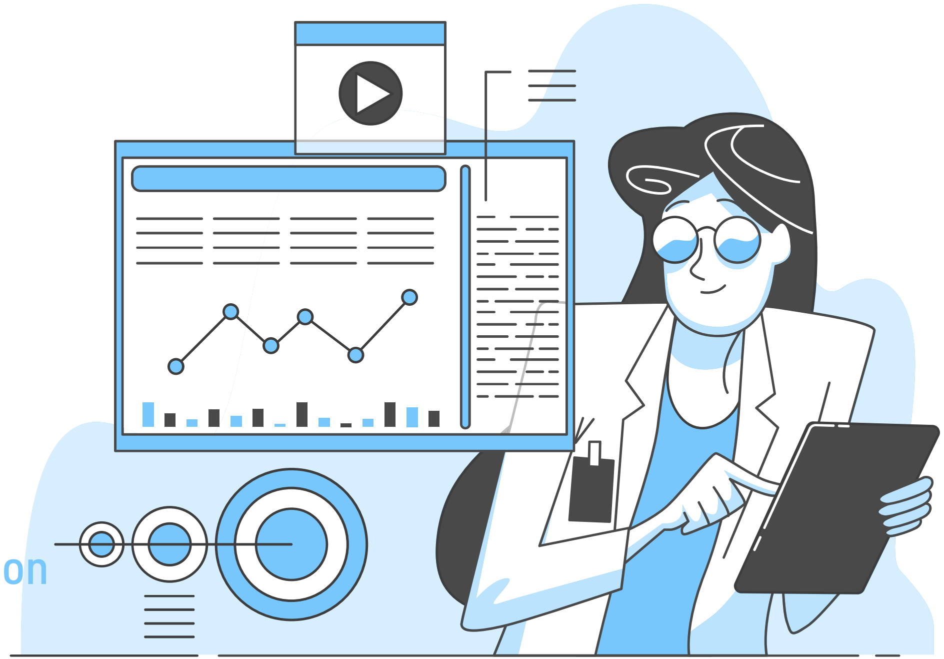
Table of Contents

01 Understanding business problem & Thinking flow

02 Source of data & Data preparation

03 Analytic Process: EDA & Segmentation

04 Insights & Recommendation



01

Understanding business problem & Thinking flow

Business Goals

Improving operational efficiency and profits

Questions?

(1) Which products are the most popular? That is, the most visited and the most frequently purchased?

(2) Which users are the most valuable users? What kind of user characteristics do these high-value users have?

(3) Which users are the most loyal users? How to improve the consumption experience of these users and increase the amount of consumption?

(4) What is the user's spending habits? Which products are users who like to buy together? Or is there a specific purchase time sequence?

(5) Which users are most effective for promotion? Can a promotion strategy consider these factors?

...

Business Solution

01

Aanalyze and find “value users”

from a large number of electronic retail transaction data

02

Propose personalized sales services

to enhance the value user experience

Thinking Flow

EDA

Understand the overall operation of e-retailers

...

RFM Model

classify users to identify value users

...

DEEP-DIVE

value users to manage business strategies

...

Hypotheses

Time of shopping

Usually at evening (8-10 p.m.), at weekend or end months of year

Unit price

Mostly from $10 -$20

Best seller products

Books, souvenirs, fashion items

02

Source of data
& Data preparation

Source of data

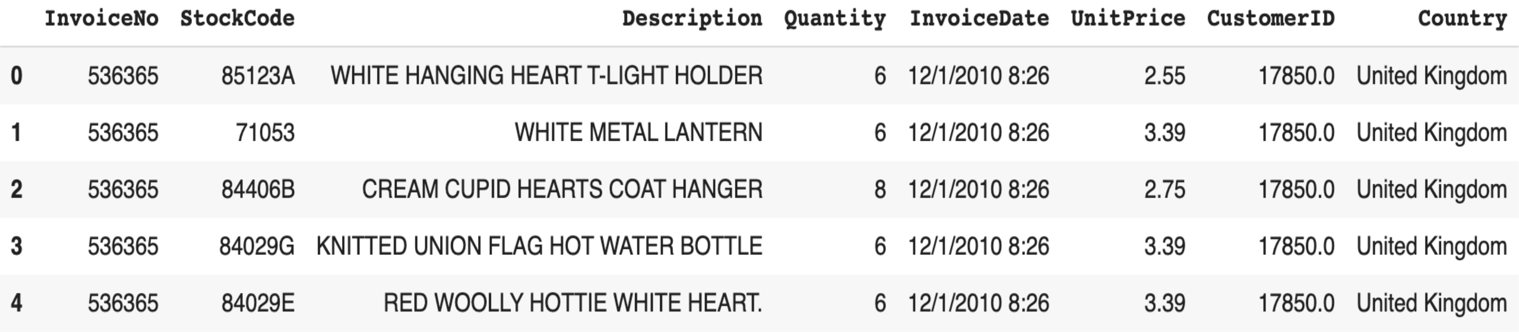
Source: Dr Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University,
London SE1 0AA, UK.

Data Set Information

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.
The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Attribute Information

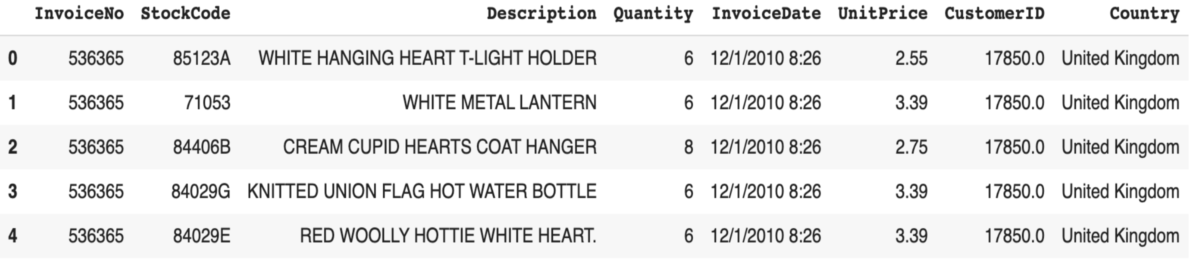
- **InvoiceNo:** Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- **Description:** Product (item) name. Nominal.
- **Quantity:** The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate:** Invice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice**: Unit price. Numeric, Product price per unit in sterling.
- **CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- **Country:** Country name. Nominal, the name of the country where each customer resides.



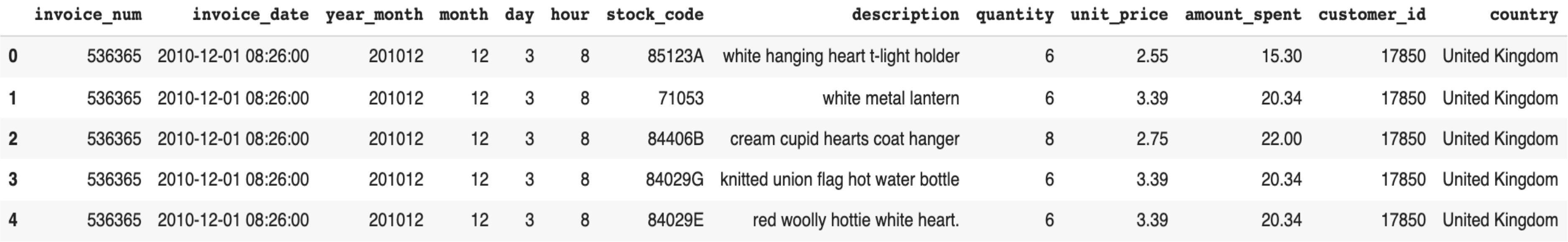
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Data Preparation

- Rename columns
- Missing data
- Duplicated Data
- Change columns type
- Reorder columns
- Add more columns



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom



	invoice_num	invoice_date	year_month	month	day	hour	stock_code	description	quantity	unit_price	amount_spent	customer_id	country
0	536365	2010-12-01 08:26:00	201012	12	3	8	85123A	white hanging heart t-light holder	6	2.55	15.30	17850	United Kingdom
1	536365	2010-12-01 08:26:00	201012	12	3	8	71053	white metal lantern	6	3.39	20.34	17850	United Kingdom
2	536365	2010-12-01 08:26:00	201012	12	3	8	84406B	cream cupid hearts coat hanger	8	2.75	22.00	17850	United Kingdom
3	536365	2010-12-01 08:26:00	201012	12	3	8	84029G	knitted union flag hot water bottle	6	3.39	20.34	17850	United Kingdom
4	536365	2010-12-01 08:26:00	201012	12	3	8	84029E	red woolly hottie white heart.	6	3.39	20.34	17850	United Kingdom

03

Analytic Process: EDA & Segmentation

Analytic Process

EDA

Understand the overall operation of e-retailers

...

RFM Model

classify users to identify value users

...

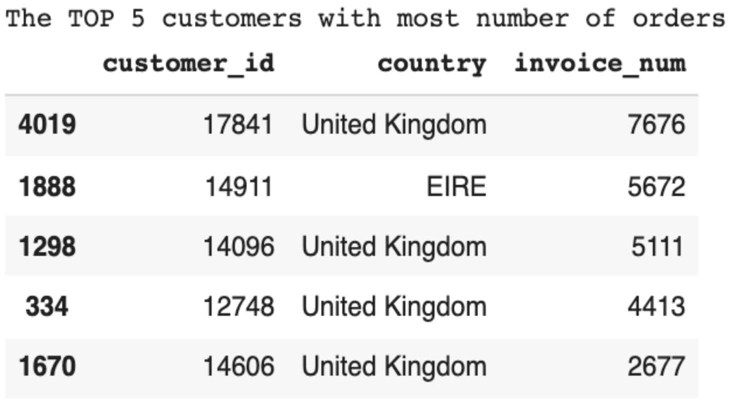
DEEP-DIVE

value users to manage business strategies

...

Explore Data Analysis
EDA

EDA

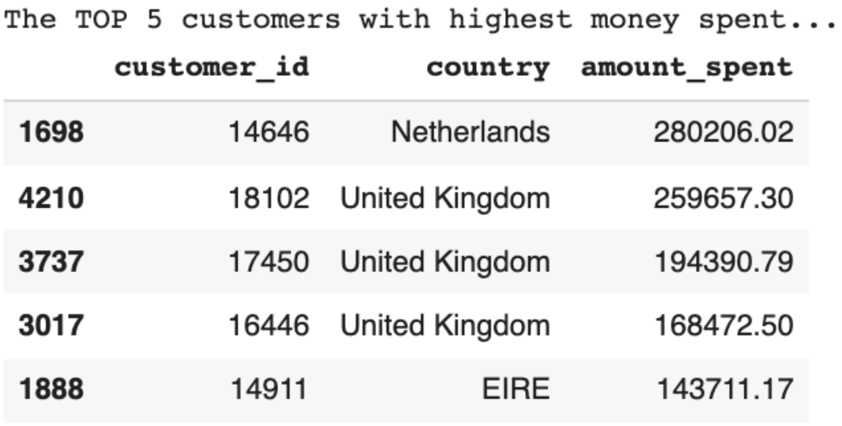


The TOP 5 customers with most number of orders

	customer_id	country	invoice_num
4019	17841	United Kingdom	7676
1888	14911	EIRE	5672
1298	14096	United Kingdom	5111
334	12748	United Kingdom	4413
1670	14606	United Kingdom	2677

- The customer with the highest number of orders comes from the United Kingdom (UK)(since it is a UK-based company).

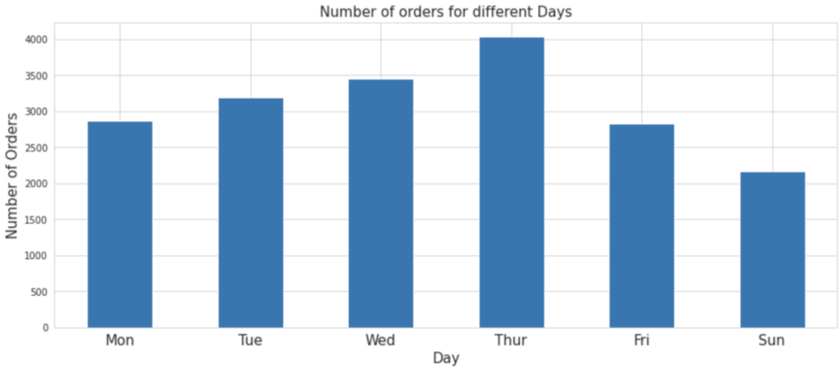
- The customer with the highest money spent on purchases comes from Netherlands



The TOP 5 customers with highest money spent...

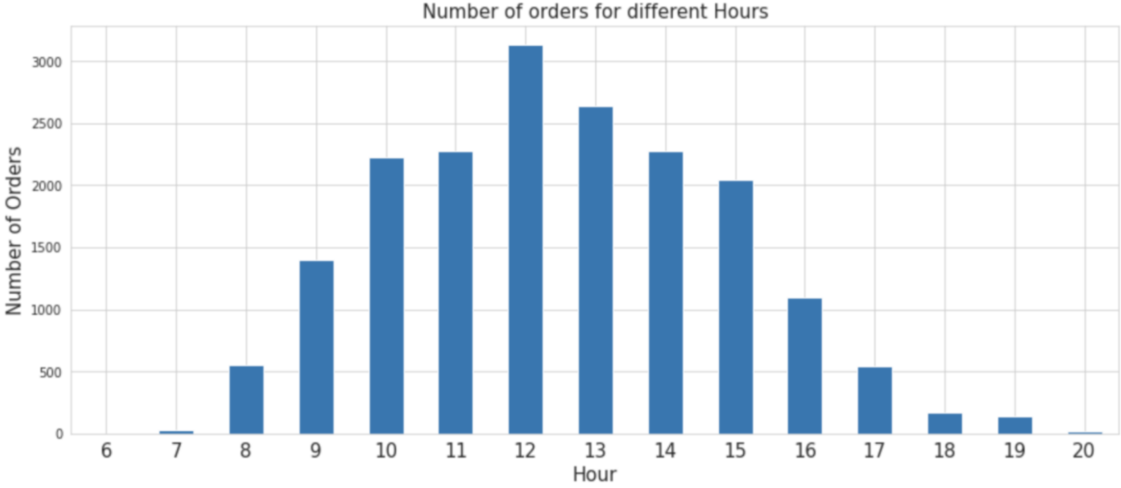
	customer_id	country	amount_spent
1698	14646	Netherlands	280206.02
4210	18102	United Kingdom	259657.30
3737	17450	United Kingdom	194390.79
3017	16446	United Kingdom	168472.50
1888	14911	EIRE	143711.17

EDA

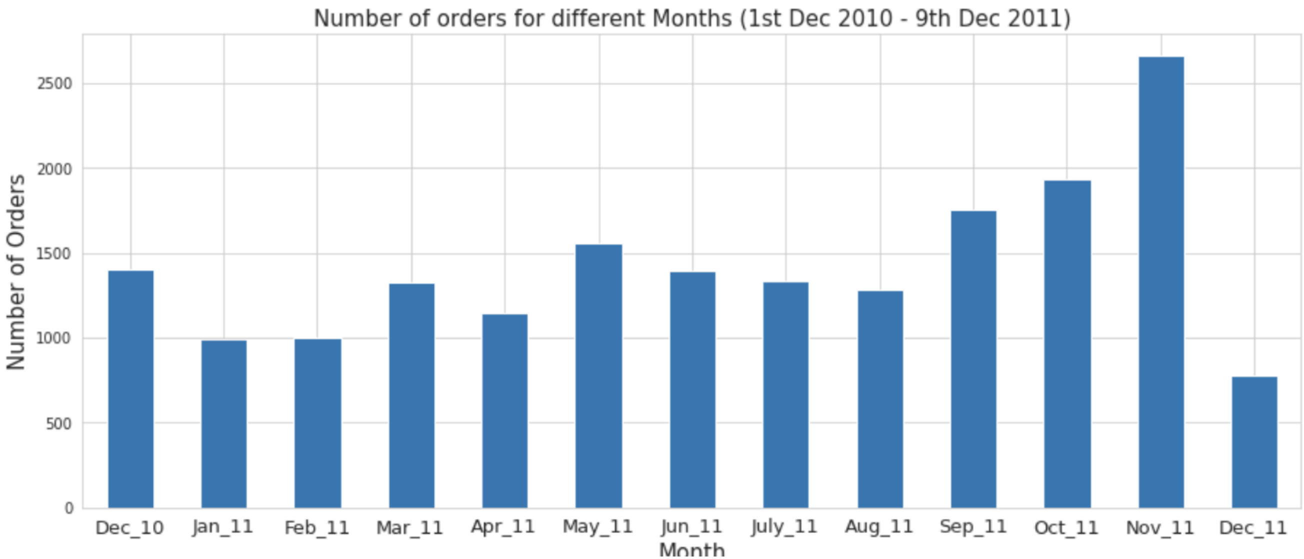


Thursday has highest number of order with 4033 orders, no order in Saturday (not at the weekend as hypotheses)

User purchase mostly from 11 a.m. to 2 p.m. (not at the evening as hypotheses)

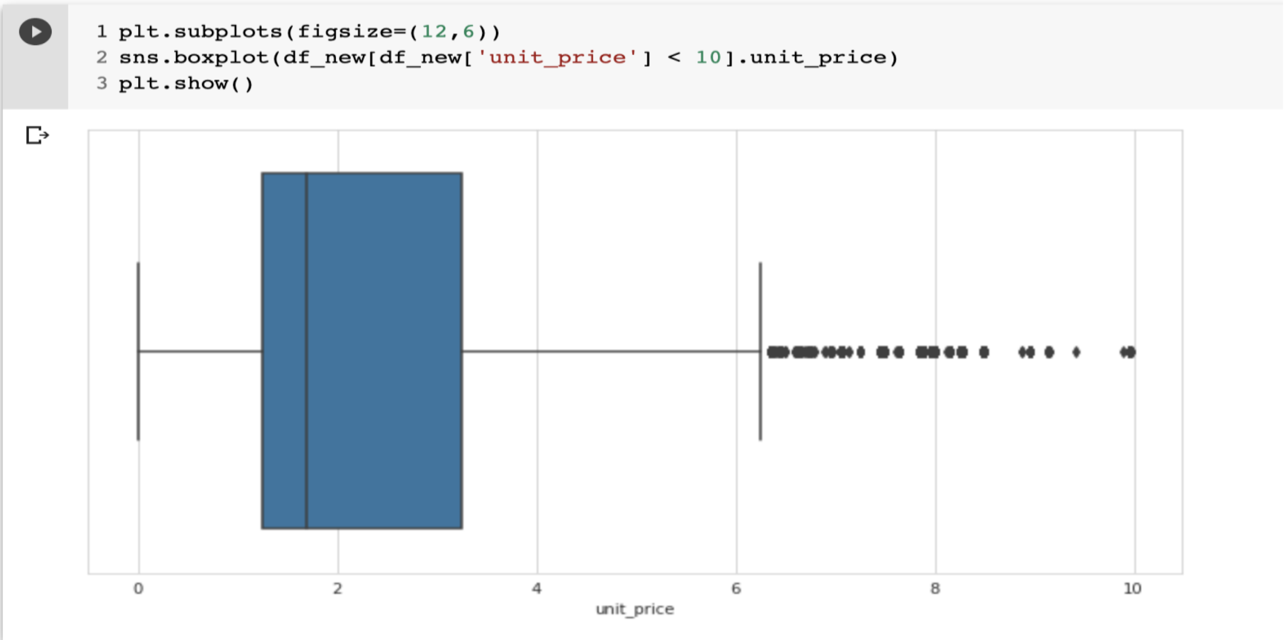


EDA



User purchase highest in the Nov-11, tends to increases from Sep -11 to Nov-11 and decreases afterward *(not high in end months of year as hypothesis)*

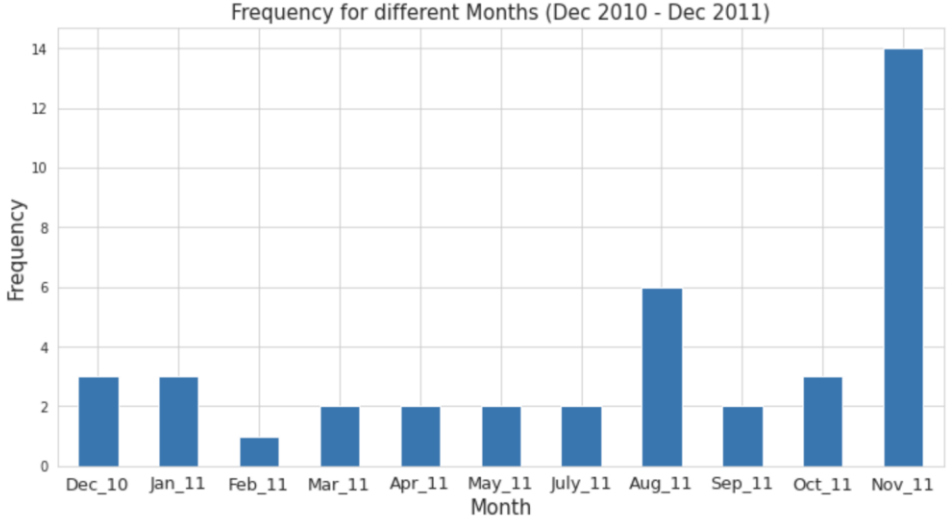
EDA

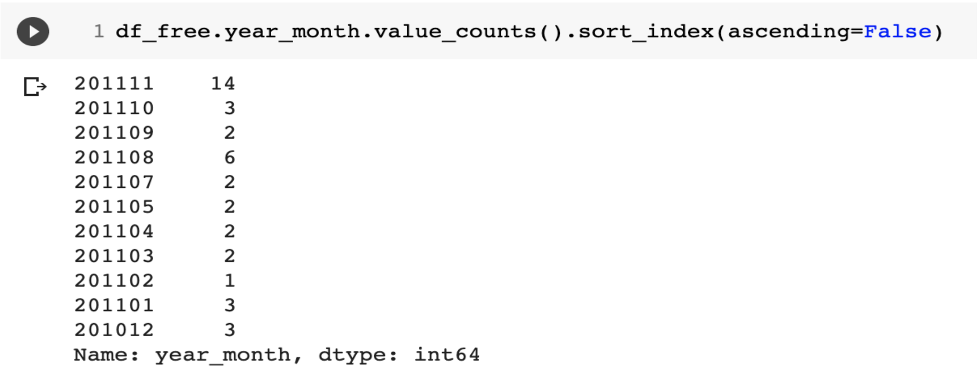


```
plt.subplots(figsize=(12,6))
sns.boxplot(df_new[df_new['unit_price'] < 10].unit_price)
plt.show()
```

Unit price mostly from $1- $4

EDA





```
df_free.year_month.value_counts().sort_index(ascending=False)
```

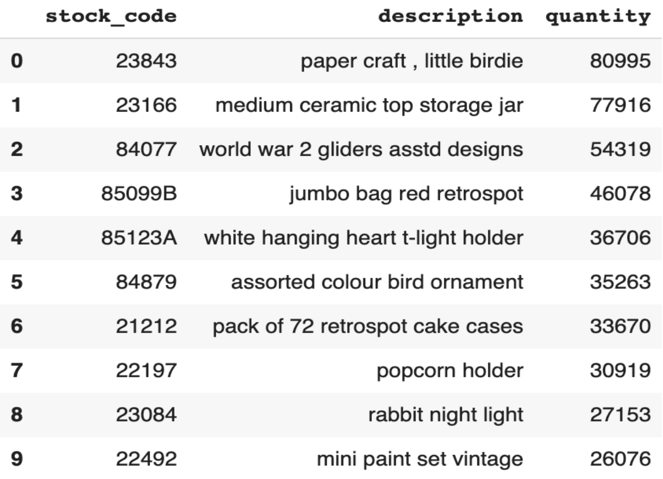
```
201111    14
201110     3
201109     2
201108     6
201107     2
201105     2
201104     2
201103     2
201102     1
201101     3
201012     3
Name: year_month, dtype: int64
```

On average, we see that the companies give 2 items for free each month.

No free items were given on June 2011 and Sept 2011.

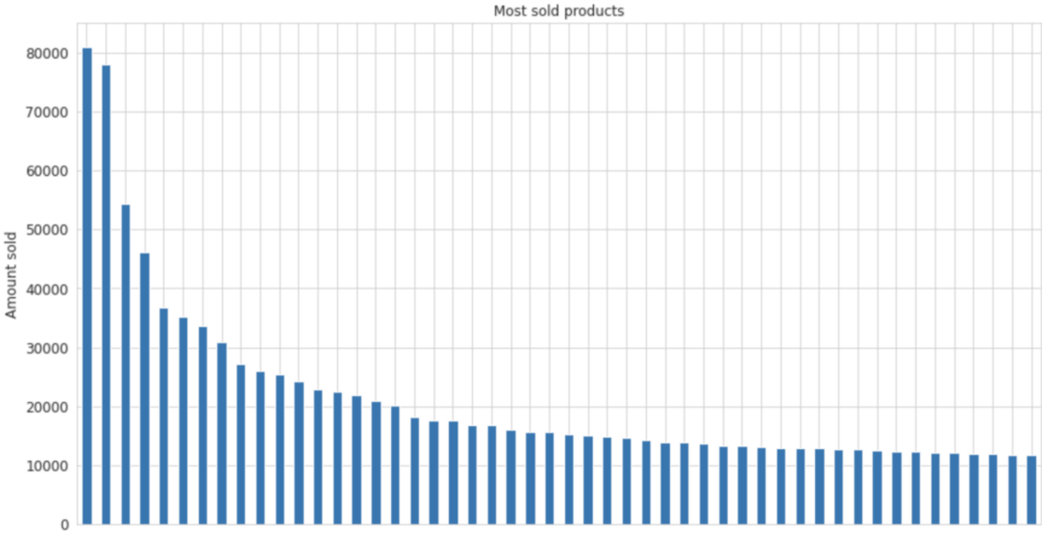
Nov-11 has highest free items

EDA



	stock_code	description	quantity
0	23843	paper craft , little birdie	80995
1	23166	medium ceramic top storage jar	77916
2	84077	world war 2 gliders asstd designs	54319
3	85099B	jumbo bag red retrospot	46078
4	85123A	white hanging heart t-light holder	36706
5	84879	assorted colour bird ornament	35263
6	21212	pack of 72 retrospot cake cases	33670
7	22197	popcorn holder	30919
8	23084	rabbit night light	27153
9	22492	mini paint set vintage	26076

- The best seller product is book and home décor, household.



Most sold products

RFM Model

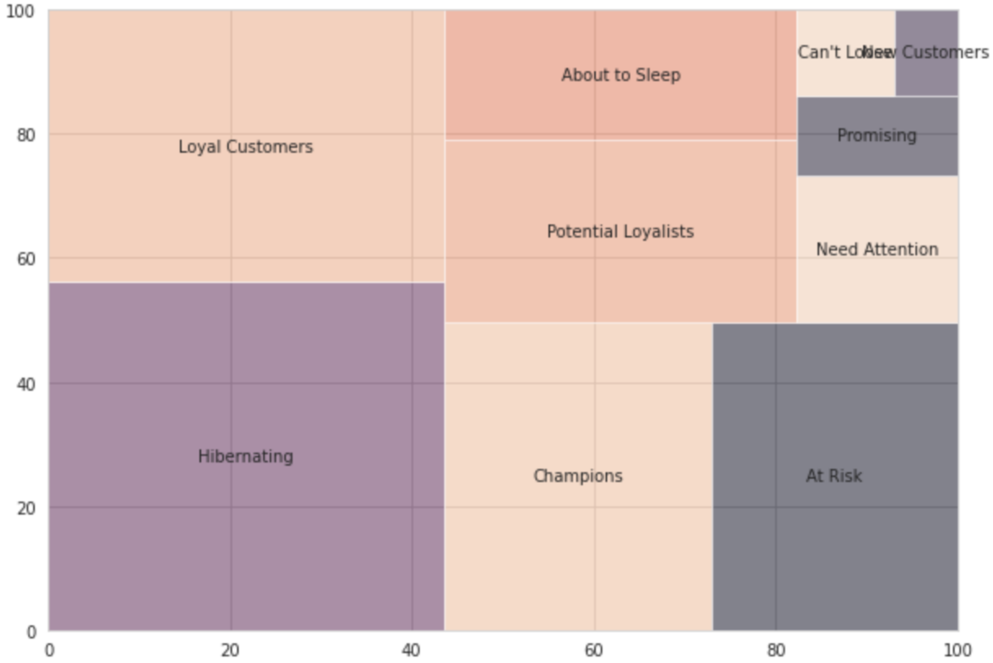
RFM Model

- RFM stands for Recency, Frequency, and Monetary value
- These RFM metrics are important indicators of a customer’s behavior because frequency and monetary value affects a customer’s lifetime value, and recency effects retention, a measure of engagement.
- RFM factors illustrate these facts:

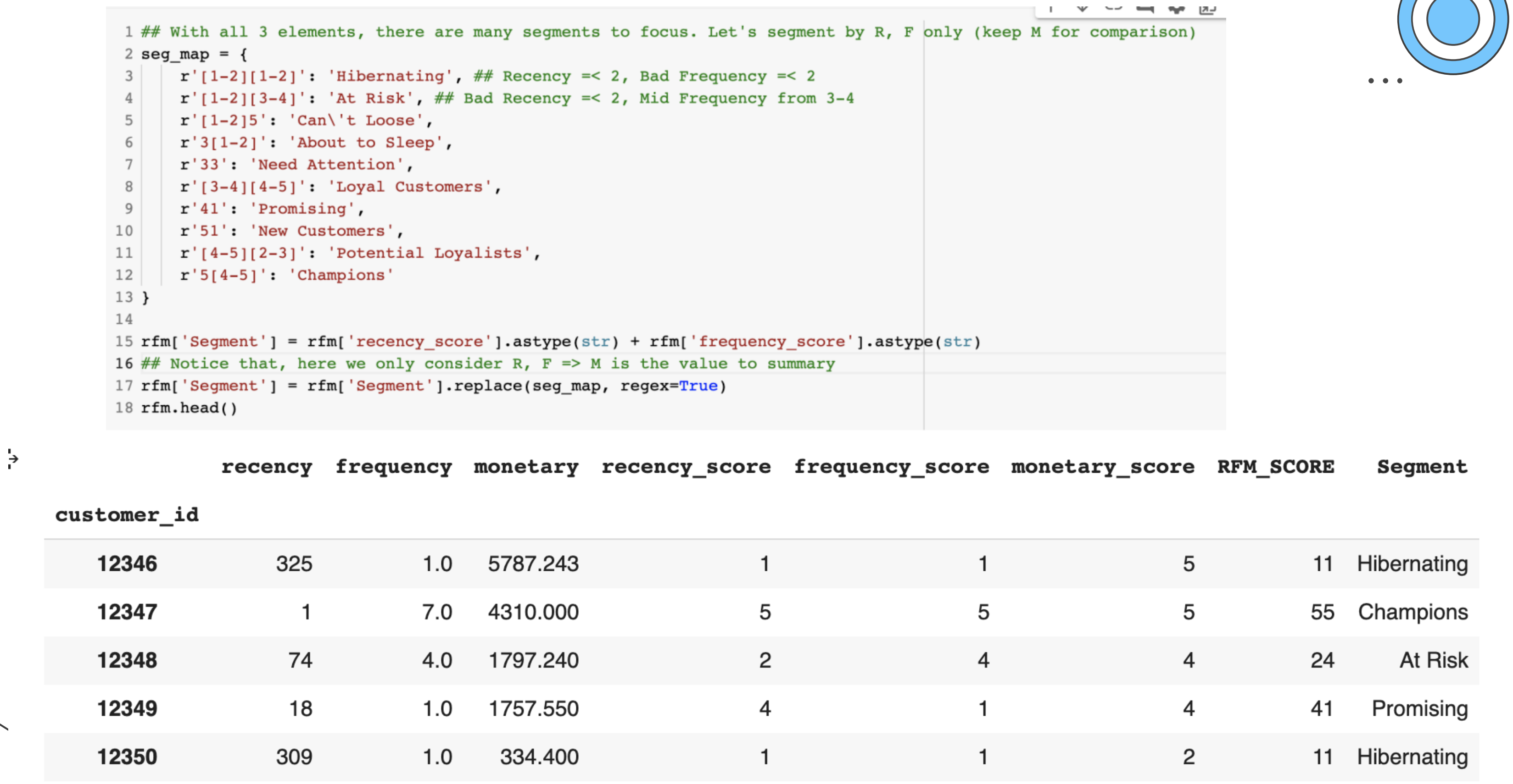
○ The more recent the purchase, the more responsive the customer is to promotions

○ The more frequently the customer buys, the more engaged and satisfied they are

○ Monetary value differentiates heavy spenders from low-value purchasers



RFM Model

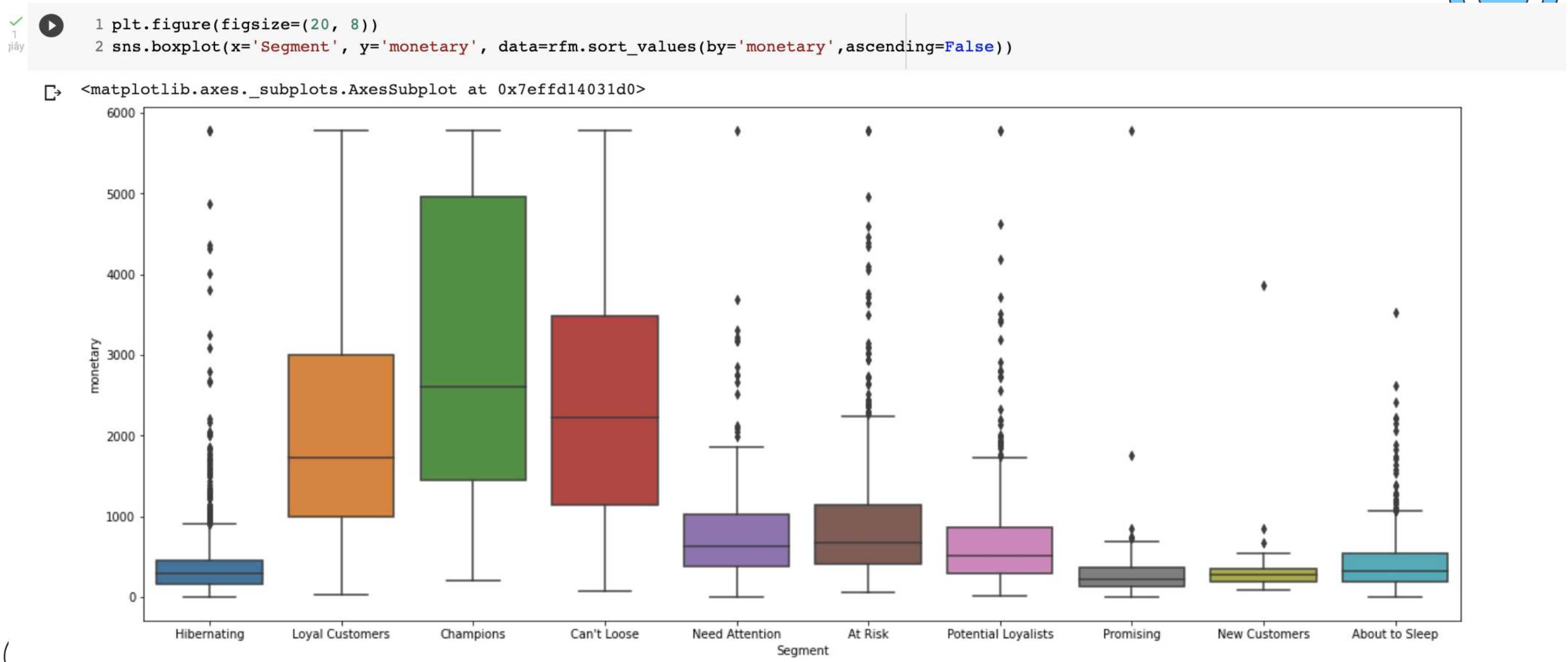


```
## With all 3 elements, there are many segments to focus. Let's segment by R, F only (keep M for comparison)
seg_map = {
    r'[1-2][1-2]': 'Hibernating', ## Recency =< 2, Bad Frequency =< 2
    r'[1-2][3-4]': 'At Risk', ## Bad Recency =< 2, Mid Frequency from 3-4
    r'[1-2]5': 'Can\'t Loose',
    r'3[1-2]': 'About to Sleep',
    r'33': 'Need Attention',
    r'[3-4][4-5]': 'Loyal Customers',
    r'41': 'Promising',
    r'51': 'New Customers',
    r'[4-5][2-3]': 'Potential Loyalists',
    r'5[4-5]': 'Champions'
}

rfm['Segment'] = rfm['recency_score'].astype(str) + rfm['frequency_score'].astype(str)
## Notice that, here we only consider R, F => M is the value to summary
rfm['Segment'] = rfm['Segment'].replace(seg_map, regex=True)
rfm.head()
```

	recency	frequency	monetary	recency_score	frequency_score	monetary_score	RFM_SCORE	Segment
customer_id								
12346	325	1.0	5787.243	1	1	5	11	Hibernating
12347	1	7.0	4310.000	5	5	5	55	Champions
12348	74	4.0	1797.240	2	4	4	24	At Risk
12349	18	1.0	1757.550	4	1	4	41	Promising
12350	309	1.0	334.400	1	1	2	11	Hibernating

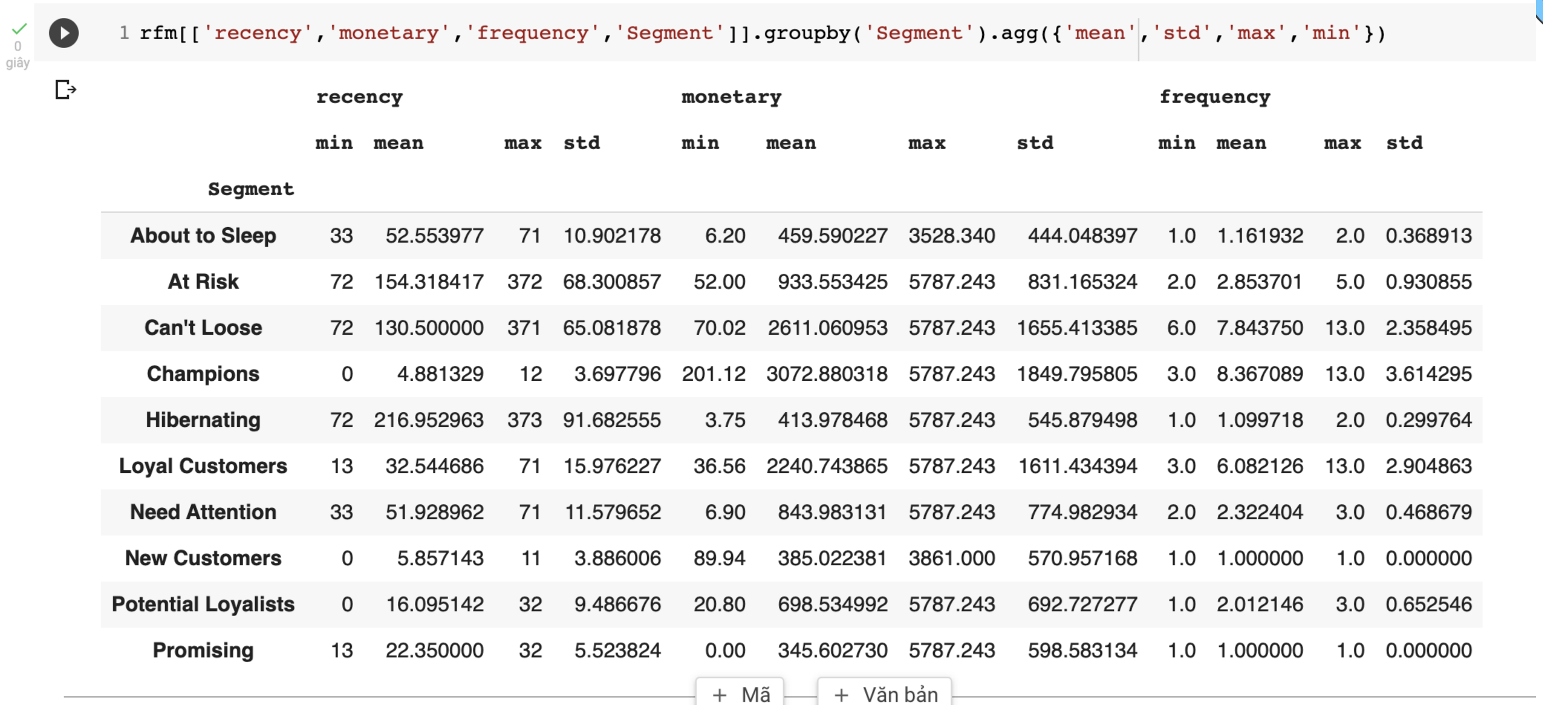
RFM Model



```
plt.figure(figsize=(20, 8))
sns.boxplot(x='Segment', y='monetary', data=rfm.sort_values(by='monetary',ascending=False))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7effd14031d0>
```

RFM Model

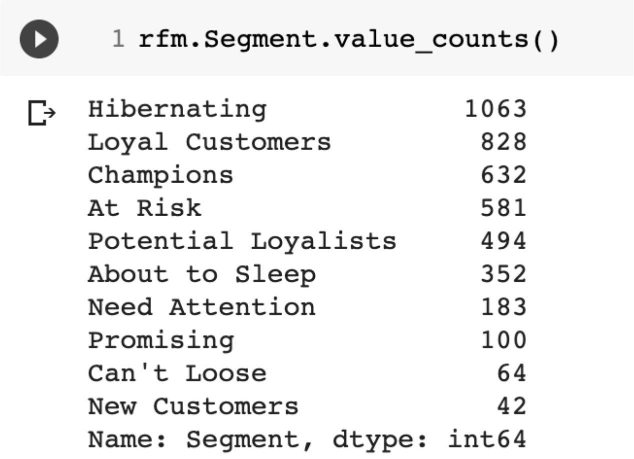


```python
rfm[['recency','monetary','frequency','Segment']].groupby('Segment').agg({'mean','std','max','min'})
```

	recency				monetary				frequency			
	min	mean	max	std	min	mean	max	std	min	mean	max	std
Segment												
About to Sleep	33	52.553977	71	10.902178	6.20	459.590227	3528.340	444.048397	1.0	1.161932	2.0	0.368913
At Risk	72	154.318417	372	68.300857	52.00	933.553425	5787.243	831.165324	2.0	2.853701	5.0	0.930855
Can't Loose	72	130.500000	371	65.081878	70.02	2611.060953	5787.243	1655.413385	6.0	7.843750	13.0	2.358495
Champions	0	4.881329	12	3.697796	201.12	3072.880318	5787.243	1849.795805	3.0	8.367089	13.0	3.614295
Hibernating	72	216.952963	373	91.682555	3.75	413.978468	5787.243	545.879498	1.0	1.099718	2.0	0.299764
Loyal Customers	13	32.544686	71	15.976227	36.56	2240.743865	5787.243	1611.434394	3.0	6.082126	13.0	2.904863
Need Attention	33	51.928962	71	11.579652	6.90	843.983131	5787.243	774.982934	2.0	2.322404	3.0	0.468679
New Customers	0	5.857143	11	3.886006	89.94	385.022381	3861.000	570.957168	1.0	1.000000	1.0	0.000000
Potential Loyalists	0	16.095142	32	9.486676	20.80	698.534992	5787.243	692.727277	1.0	2.012146	3.0	0.652546
Promising	13	22.350000	32	5.523824	0.00	345.602730	5787.243	598.583134	1.0	1.000000	1.0	0.000000

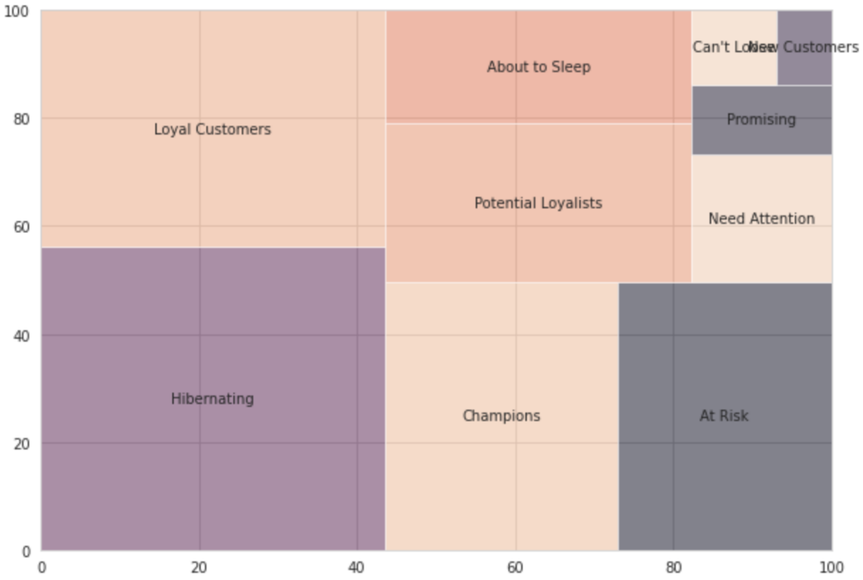
Segmentation Analysis

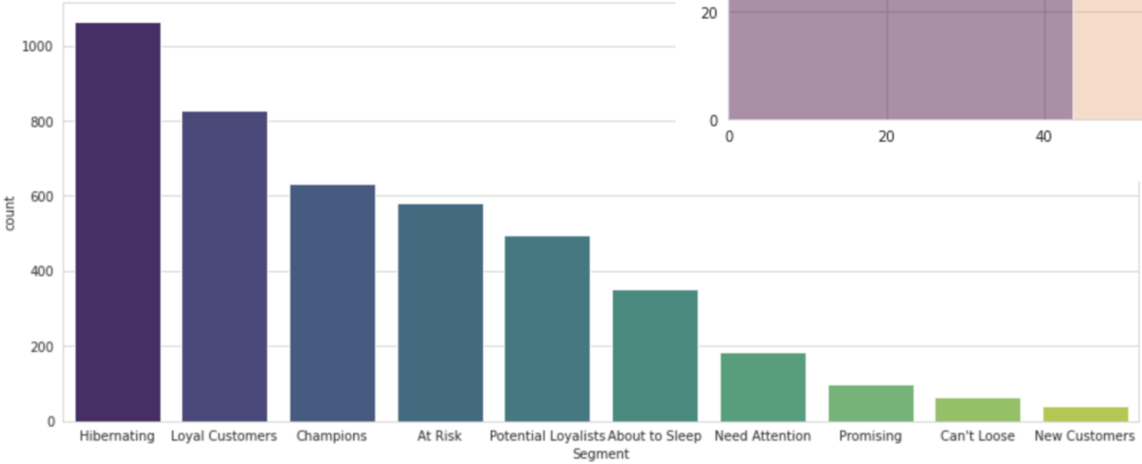
RFM Model



```
rfm.Segment.value_counts()
```

```
Hibernating           1063
Loyal Customers        828
Champions              632
At Risk                581
Potential Loyalists    494
About to Sleep         352
Need Attention         183
Promising              100
Can't Loose             64
New Customers           42
Name: Segment, dtype: int64
```

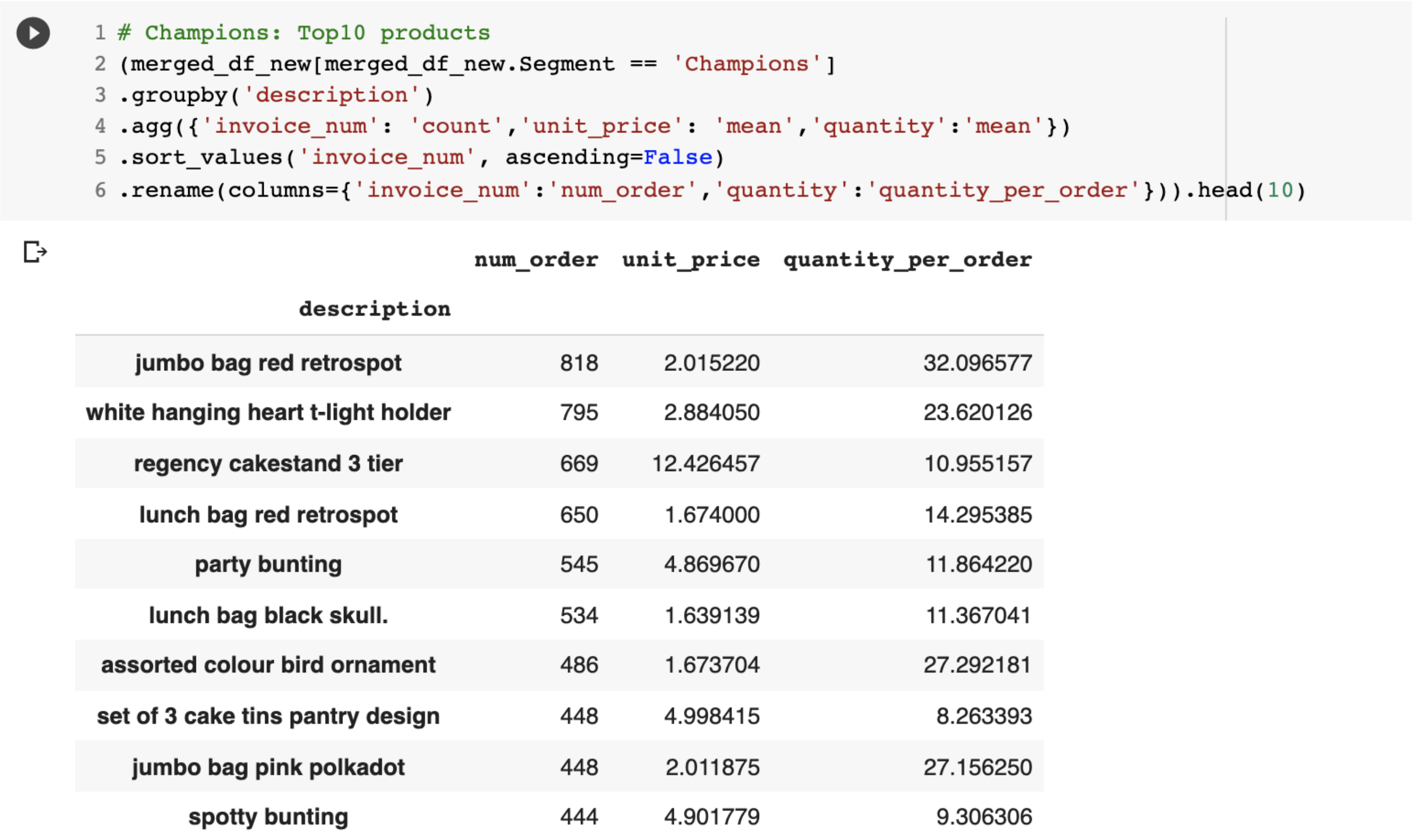




Segmentation Analysis

Deep-dive Analysis “Value Users”

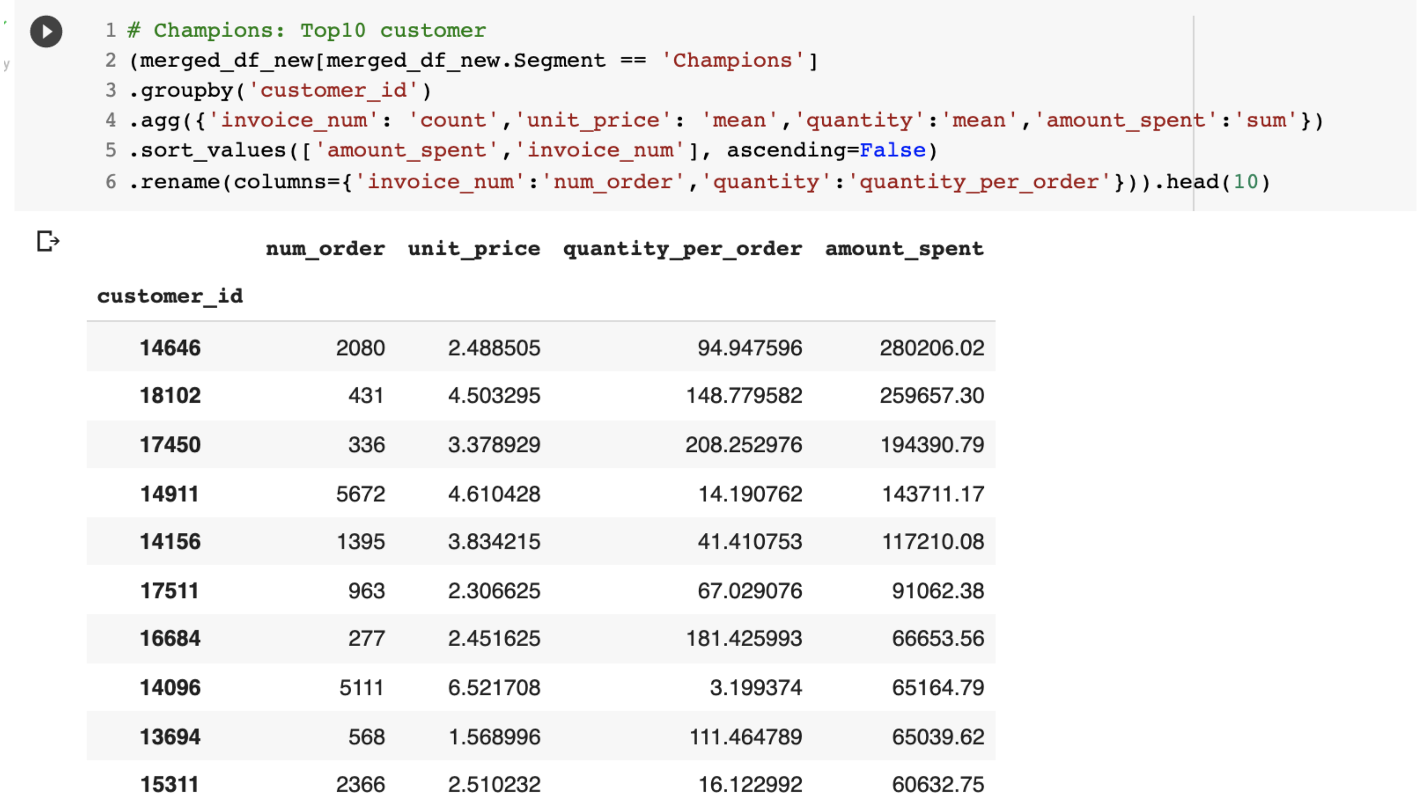
Champions segment



```python
# Champions: Top10 products
(merged_df_new[merged_df_new.Segment == 'Champions']
.groupby('description')
.agg({'invoice_num': 'count','unit_price': 'mean','quantity':'mean'})
.sort_values('invoice_num', ascending=False)
.rename(columns={'invoice_num':'num_order','quantity':'quantity_per_order'})).head(10)
```

	num_order	unit_price	quantity_per_order
description			
jumbo bag red retrospot	818	2.015220	32.096577
white hanging heart t-light holder	795	2.884050	23.620126
regency cakestand 3 tier	669	12.426457	10.955157
lunch bag red retrospot	650	1.674000	14.295385
party bunting	545	4.869670	11.864220
lunch bag black skull.	534	1.639139	11.367041
assorted colour bird ornament	486	1.673704	27.292181
set of 3 cake tins pantry design	448	4.998415	8.263393
jumbo bag pink polkadot	448	2.011875	27.156250
spotty bunting	444	4.901779	9.306306

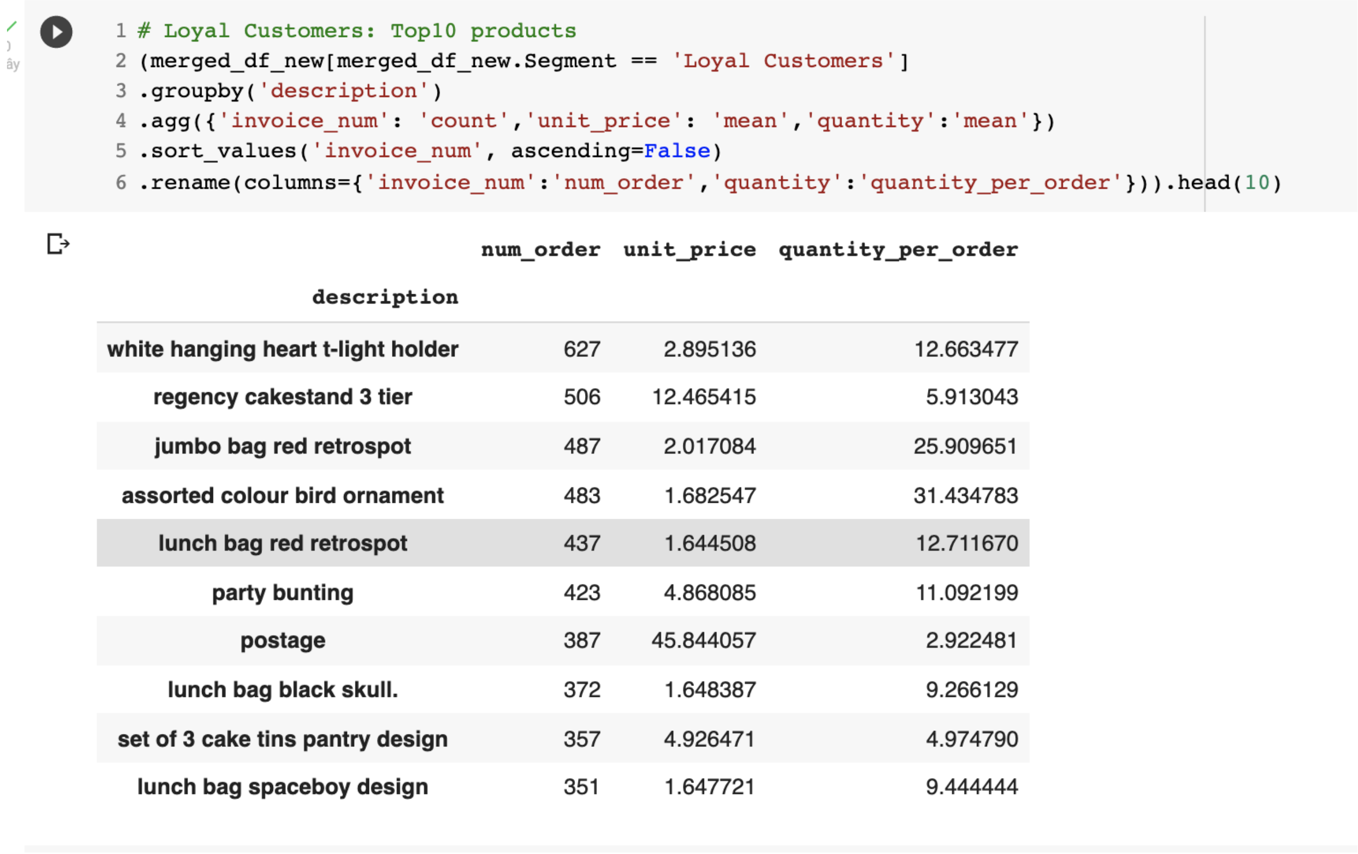
Champions segment



```python
# Champions: Top10 customer
(merged_df_new[merged_df_new.Segment == 'Champions']
.groupby('customer_id')
.agg({'invoice_num': 'count','unit_price': 'mean','quantity':'mean','amount_spent':'sum'})
.sort_values(['amount_spent','invoice_num'], ascending=False)
.rename(columns={'invoice_num':'num_order','quantity':'quantity_per_order'})).head(10)
```

	num_order	unit_price	quantity_per_order	amount_spent
customer_id				
14646	2080	2.488505	94.947596	280206.02
18102	431	4.503295	148.779582	259657.30
17450	336	3.378929	208.252976	194390.79
14911	5672	4.610428	14.190762	143711.17
14156	1395	3.834215	41.410753	117210.08
17511	963	2.306625	67.029076	91062.38
16684	277	2.451625	181.425993	66653.56
14096	5111	6.521708	3.199374	65164.79
13694	568	1.568996	111.464789	65039.62
15311	2366	2.510232	16.122992	60632.75

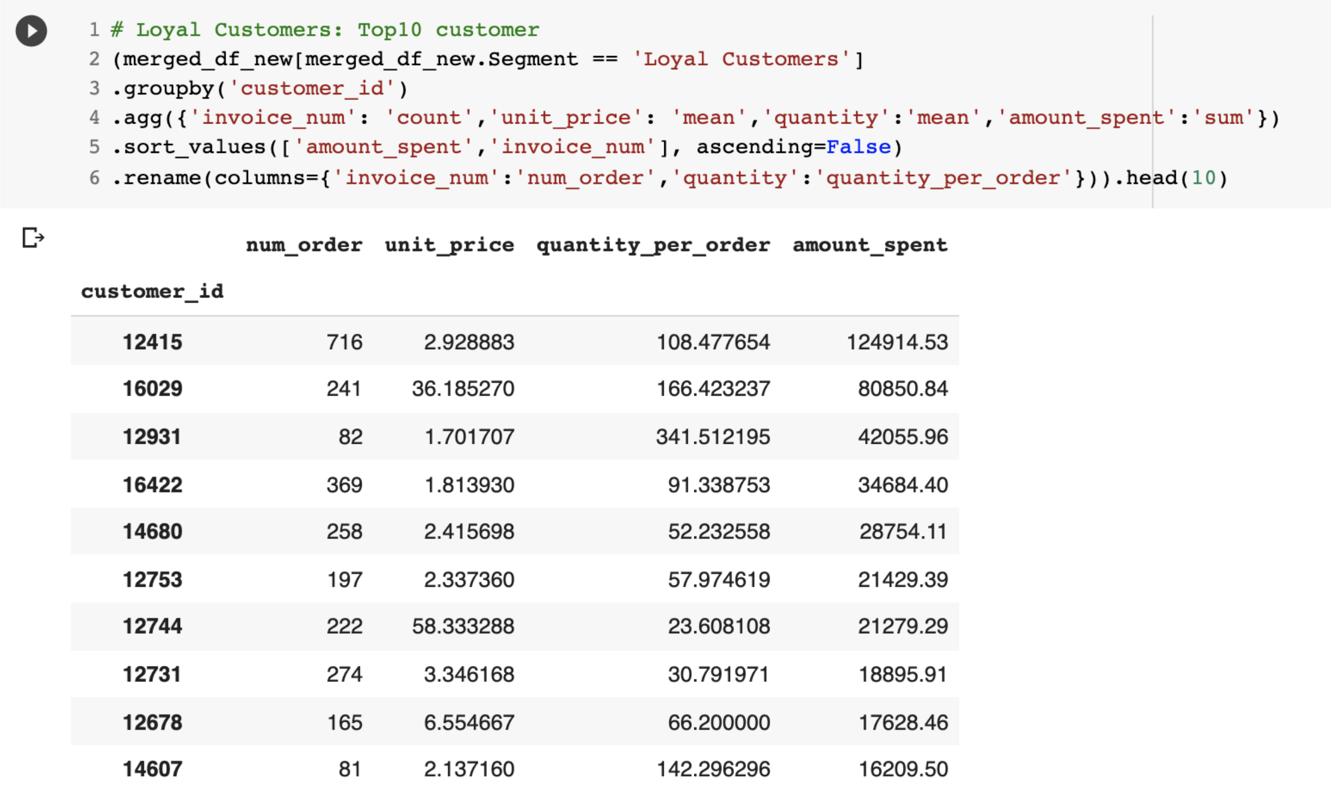
Loyal Customers Segment



```python
# Loyal Customers: Top10 products
(merged_df_new[merged_df_new.Segment == 'Loyal Customers']
.groupby('description')
.agg({'invoice_num': 'count','unit_price': 'mean','quantity':'mean'})
.sort_values('invoice_num', ascending=False)
.rename(columns={'invoice_num':'num_order','quantity':'quantity_per_order'})).head(10)
```

	num_order	unit_price	quantity_per_order
description			
white hanging heart t-light holder	627	2.895136	12.663477
regency cakestand 3 tier	506	12.465415	5.913043
jumbo bag red retrospot	487	2.017084	25.909651
assorted colour bird ornament	483	1.682547	31.434783
lunch bag red retrospot	437	1.644508	12.711670
party bunting	423	4.868085	11.092199
postage	387	45.844057	2.922481
lunch bag black skull.	372	1.648387	9.266129
set of 3 cake tins pantry design	357	4.926471	4.974790
lunch bag spaceboy design	351	1.647721	9.444444

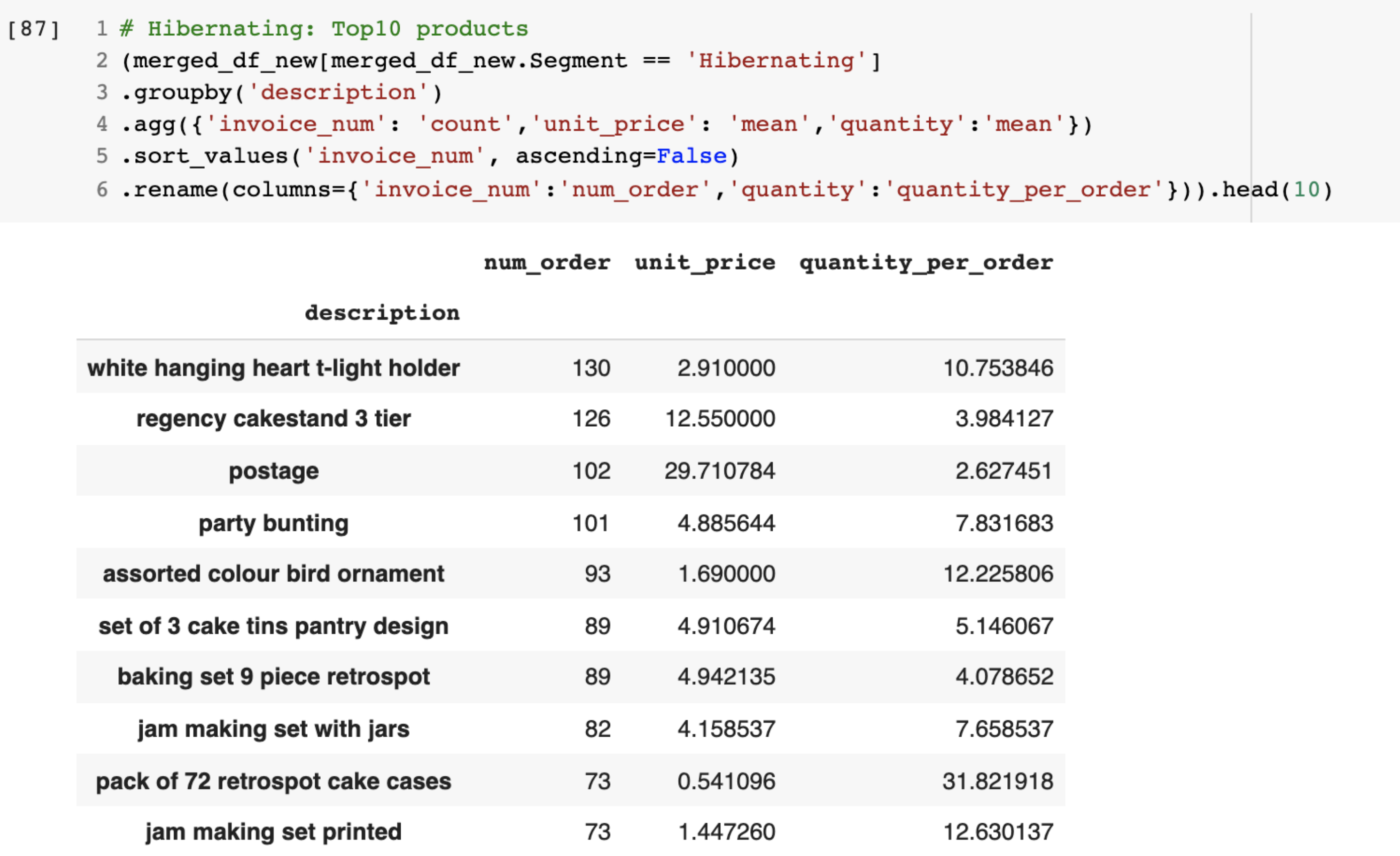
Loyal Customers Segment



```
# Loyal Customers: Top10 customer
(merged_df_new[merged_df_new.Segment == 'Loyal Customers']
.groupby('customer_id')
.agg({'invoice_num': 'count','unit_price': 'mean','quantity':'mean','amount_spent':'sum'})
.sort_values(['amount_spent','invoice_num'], ascending=False)
.rename(columns={'invoice_num':'num_order','quantity':'quantity_per_order'})).head(10)
```

	num_order	unit_price	quantity_per_order	amount_spent
customer_id				
12415	716	2.928883	108.477654	124914.53
16029	241	36.185270	166.423237	80850.84
12931	82	1.701707	341.512195	42055.96
16422	369	1.813930	91.338753	34684.40
14680	258	2.415698	52.232558	28754.11
12753	197	2.337360	57.974619	21429.39
12744	222	58.333288	23.608108	21279.29
12731	274	3.346168	30.791971	18895.91
12678	165	6.554667	66.200000	17628.46
14607	81	2.137160	142.296296	16209.50

Hibernating segment

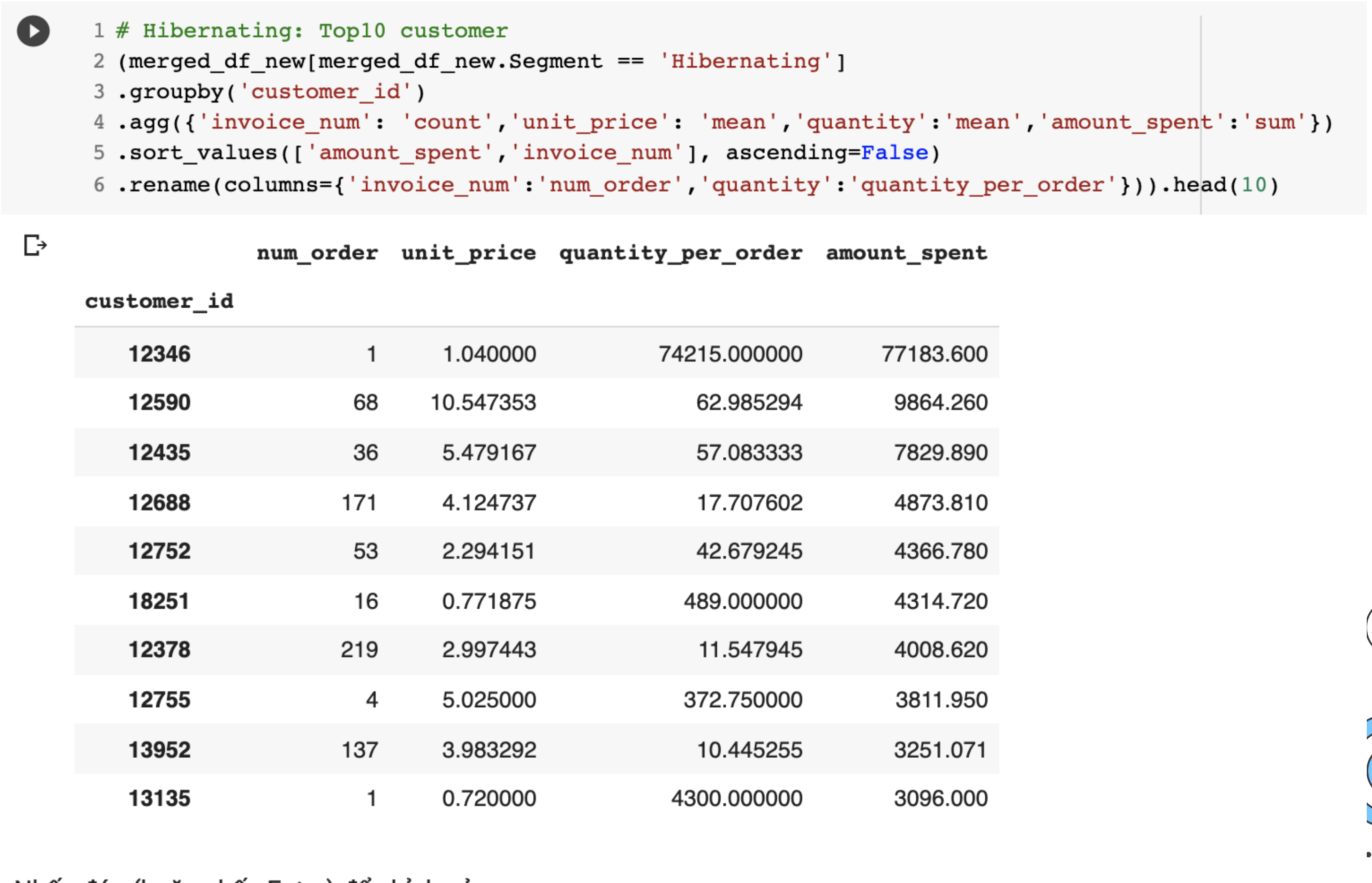


[87]

```
# Hibernating: Top10 products
(merged_df_new[merged_df_new.Segment == 'Hibernating']
.groupby('description')
.agg({'invoice_num': 'count','unit_price': 'mean','quantity':'mean'})
.sort_values('invoice_num', ascending=False)
.rename(columns={'invoice_num':'num_order','quantity':'quantity_per_order'})).head(10)
```

	num_order	unit_price	quantity_per_order
description			
white hanging heart t-light holder	130	2.910000	10.753846
regency cakestand 3 tier	126	12.550000	3.984127
postage	102	29.710784	2.627451
party bunting	101	4.885644	7.831683
assorted colour bird ornament	93	1.690000	12.225806
set of 3 cake tins pantry design	89	4.910674	5.146067
baking set 9 piece retrospot	89	4.942135	4.078652
jam making set with jars	82	4.158537	7.658537
pack of 72 retrospot cake cases	73	0.541096	31.821918
jam making set printed	73	1.447260	12.630137

Hibernating segment



```
# Hibernating: Top10 customer
(merged_df_new[merged_df_new.Segment == 'Hibernating']
.groupby('customer_id')
.agg({'invoice_num': 'count','unit_price': 'mean','quantity':'mean','amount_spent':'sum'})
.sort_values(['amount_spent','invoice_num'], ascending=False)
.rename(columns={'invoice_num':'num_order','quantity':'quantity_per_order'})).head(10)
```

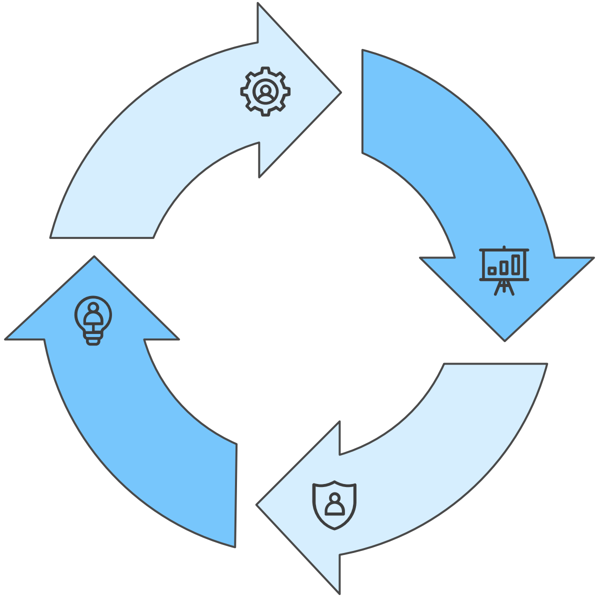
	num_order	unit_price	quantity_per_order	amount_spent
customer_id				
12346	1	1.040000	74215.000000	77183.600
12590	68	10.547353	62.985294	9864.260
12435	36	5.479167	57.083333	7829.890
12688	171	4.124737	17.707602	4873.810
12752	53	2.294151	42.679245	4366.780
18251	16	0.771875	489.000000	4314.720
12378	219	2.997443	11.547945	4008.620
12755	4	5.025000	372.750000	3811.950
13952	137	3.983292	10.445255	3251.071
13135	1	0.720000	4300.000000	3096.000

Nhấp đúp (hoặc nhấn Enter) để chỉnh sửa

04

Insights and Recommendations

Insights & Recommendations



Time

Highest purchasing at 12 p.m, on Thursday, in Nov

Price per item

Price mostly under $10

Best Seller Products

Book, home décor, house hold

Value Customers

Champions,
Loyal Customers,
Hibernating segment

Thanks!

Do you have any questions?