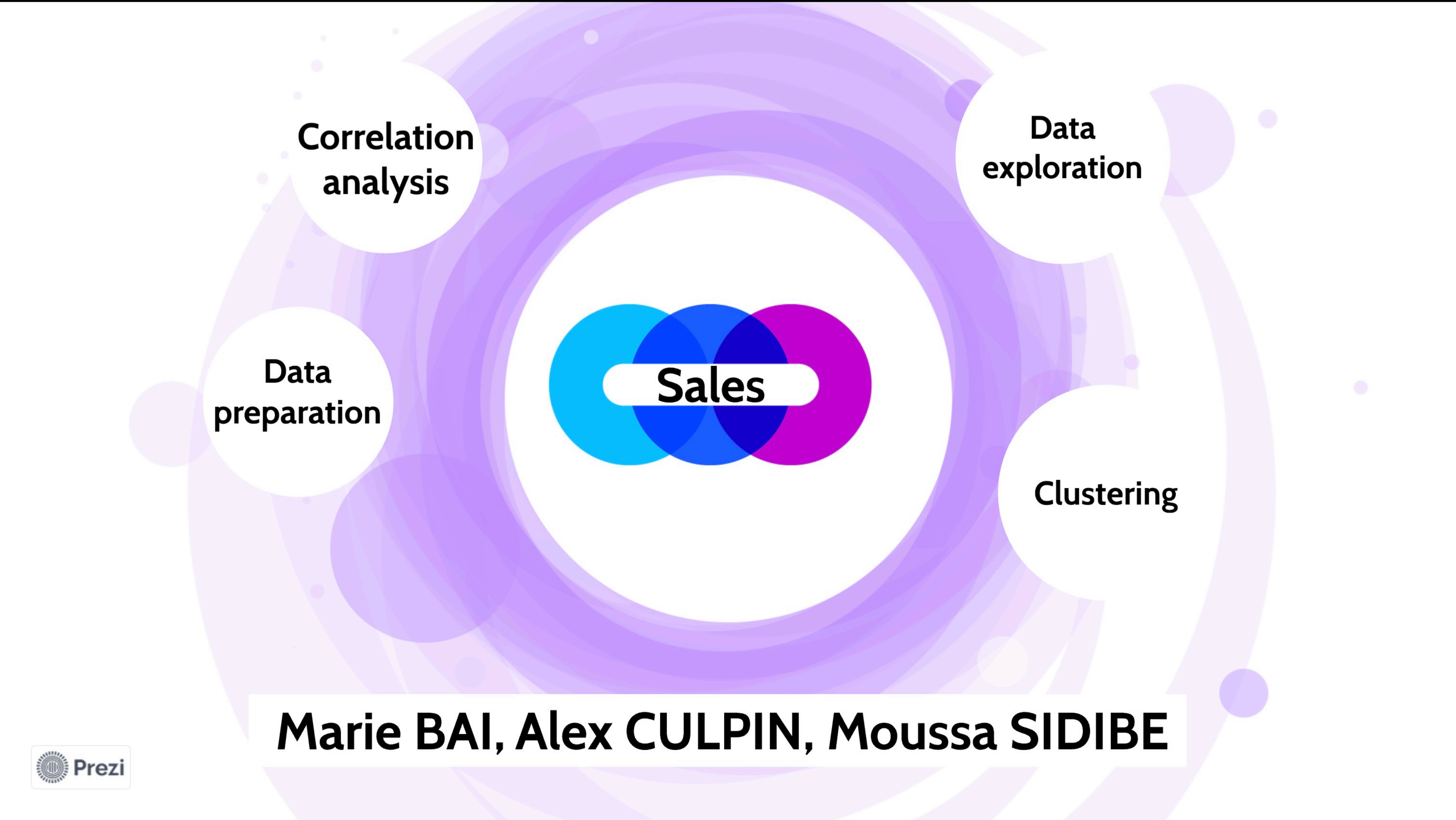


Sales

Correlation analysis

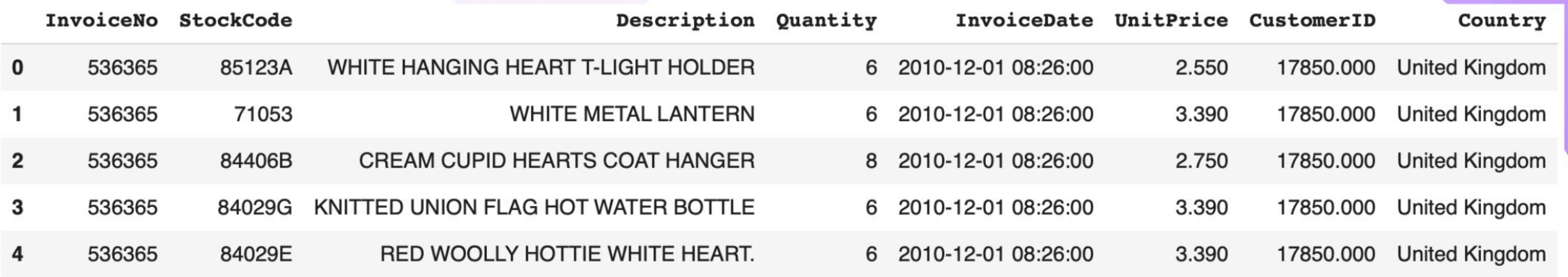
Data exploration

Data preparation

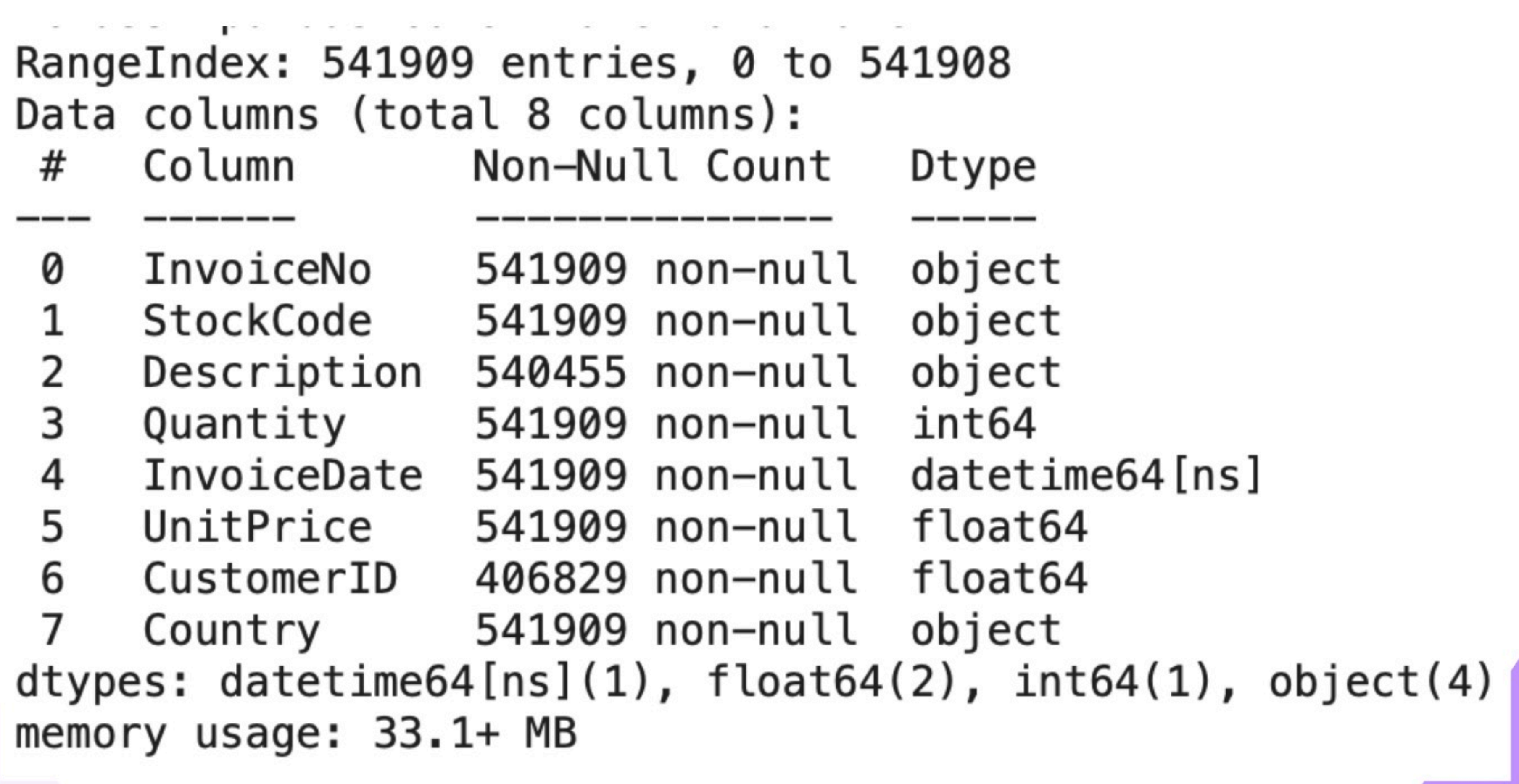
Clustering

Marie BAI, Alex CULPIN, Moussa SIDIBE

Data preparation



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.550	17850.000	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.390	17850.000	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.750	17850.000	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.390	17850.000	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.390	17850.000	United Kingdom

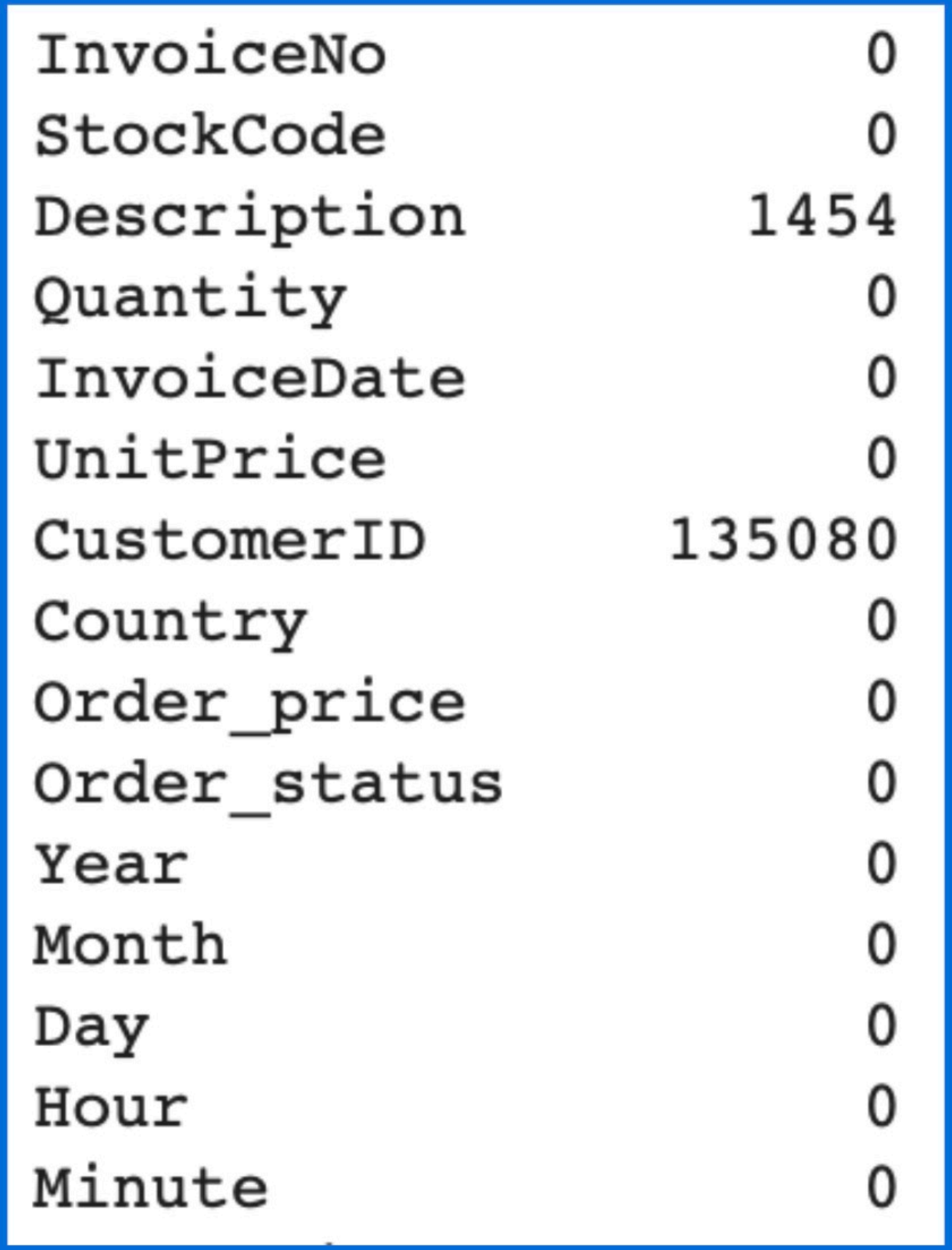


```
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

Missing values

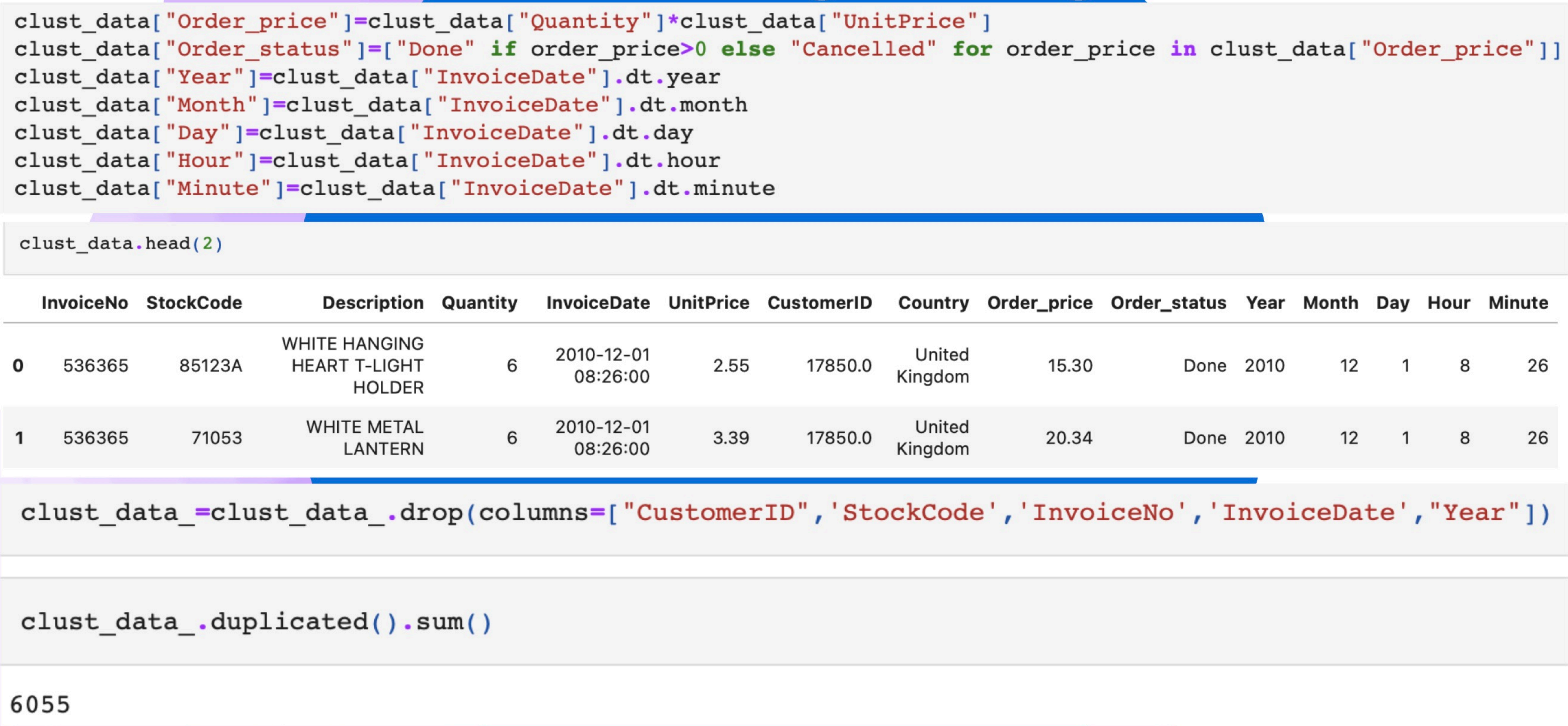
Feature engeneering

Missing values



```
InvoiceNo             0
StockCode             0
Description        1454
Quantity              0
InvoiceDate           0
UnitPrice             0
CustomerID       135080
Country               0
Order_price           0
Order_status          0
Year                  0
Month                 0
Day                   0
Hour                  0
Minute                0
```

Feature engeneering



```
clust_data["Order_price"]=clust_data["Quantity"]*clust_data["UnitPrice"]
clust_data["Order_status"]=["Done" if order_price>0 else "Cancelled" for order_price in clust_data["Order_price"]]
clust_data["Year"]=clust_data["InvoiceDate"].dt.year
clust_data["Month"]=clust_data["InvoiceDate"].dt.month
clust_data["Day"]=clust_data["InvoiceDate"].dt.day
clust_data["Hour"]=clust_data["InvoiceDate"].dt.hour
clust_data["Minute"]=clust_data["InvoiceDate"].dt.minute
```

```
clust_data.head(2)
```

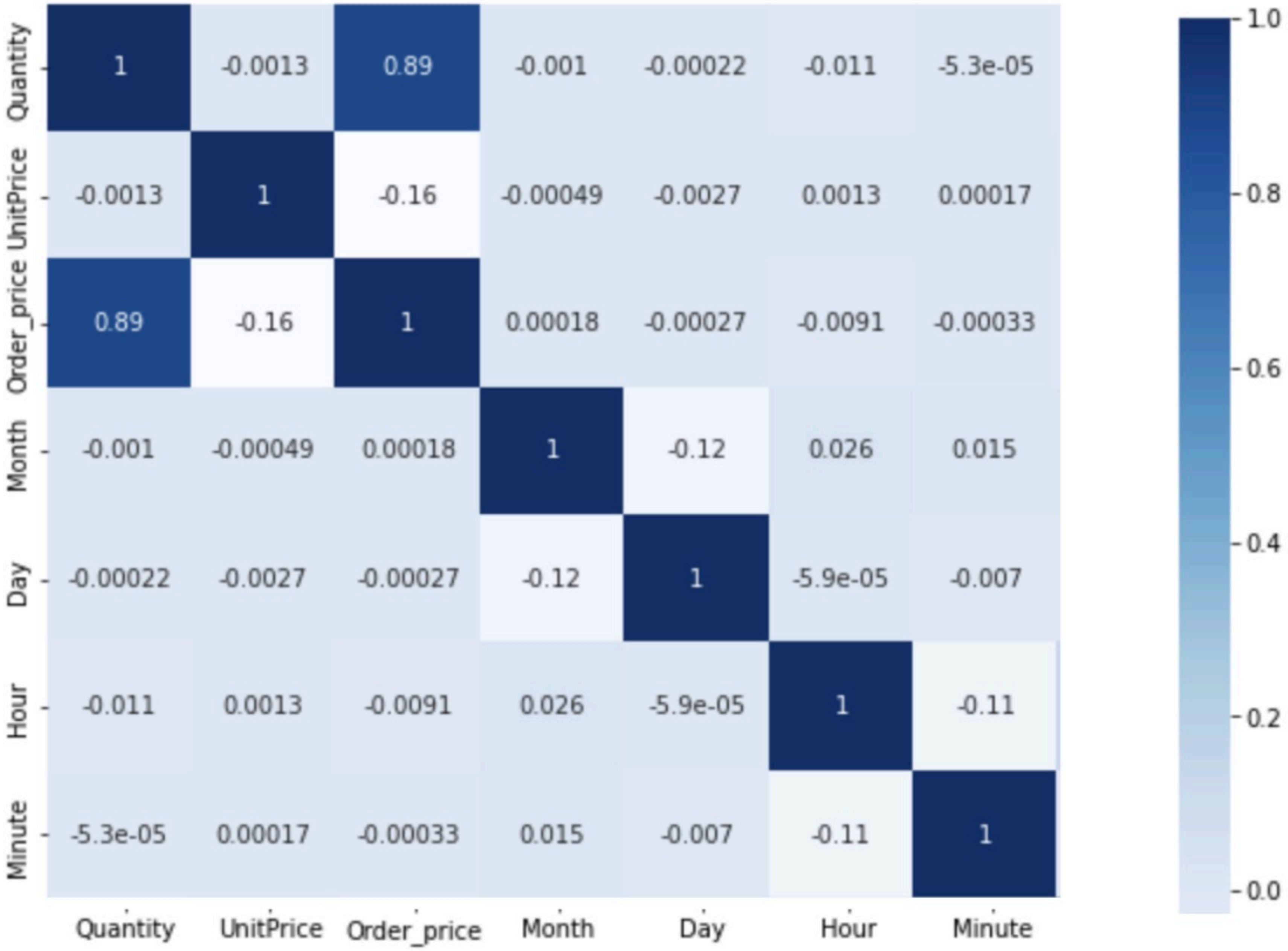
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Order_price	Order_status	Year	Month	Day	Hour	Minute
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30	Done	2010	12	1	8	26
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	Done	2010	12	1	8	26

```
clust_data_=clust_data_.drop(columns=["CustomerID",'StockCode','InvoiceNo','InvoiceDate',"Year"])
```

```
clust_data_.duplicated().sum()
```

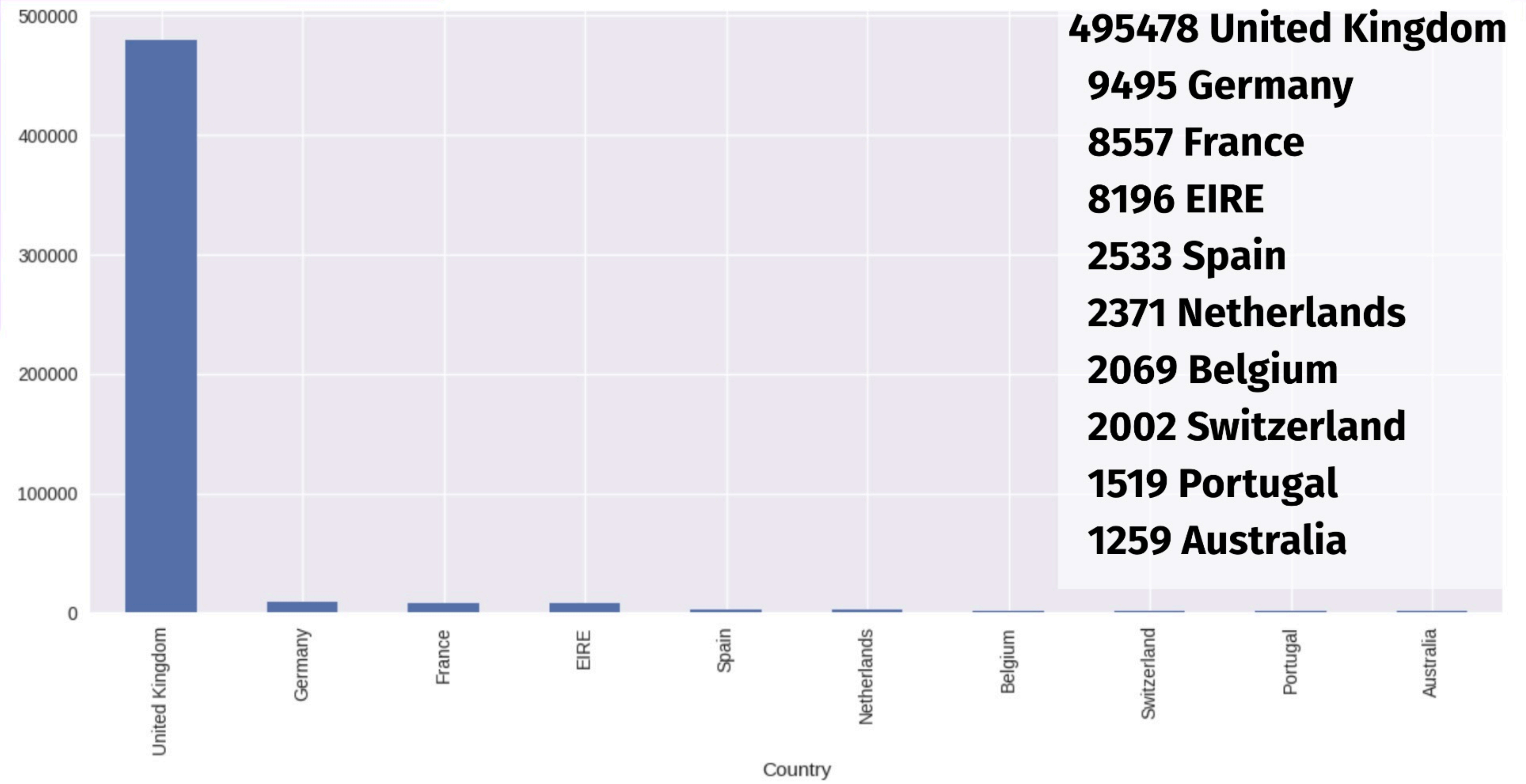
```
6055
```

Correlation



	Quantity	UnitPrice	Order_price	Month	Day	Hour	Minute
Quantity	1	-0.0013	0.89	-0.001	-0.00022	-0.011	-5.3e-05
UnitPrice	-0.0013	1	-0.16	-0.00049	-0.0027	0.0013	0.00017
Order_price	0.89	-0.16	1	0.00018	-0.00027	-0.0091	-0.00033
Month	-0.001	-0.00049	0.00018	1	-0.12	0.026	0.015
Day	-0.00022	-0.0027	-0.00027	-0.12	1	-5.9e-05	-0.007
Hour	-0.011	0.0013	-0.0091	0.026	-5.9e-05	1	-0.11
Minute	-5.3e-05	0.00017	-0.00033	0.015	-0.007	-0.11	1

Data Exploration



495478	United Kingdom
9495	Germany
8557	France
8196	EIRE
2533	Spain
2371	Netherlands
2069	Belgium
2002	Switzerland
1519	Portugal
1259	Australia

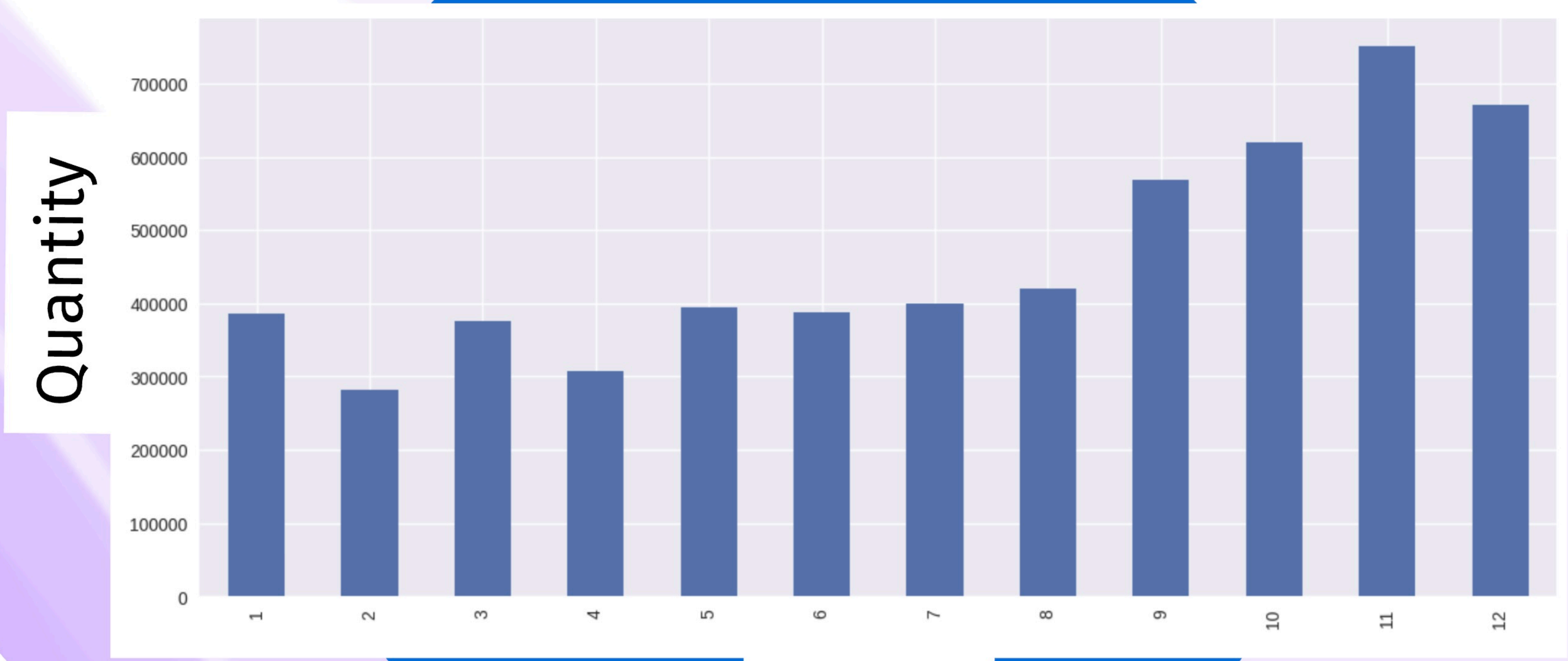
Quant/ Month

Quant/ Day

Quant/ Hour

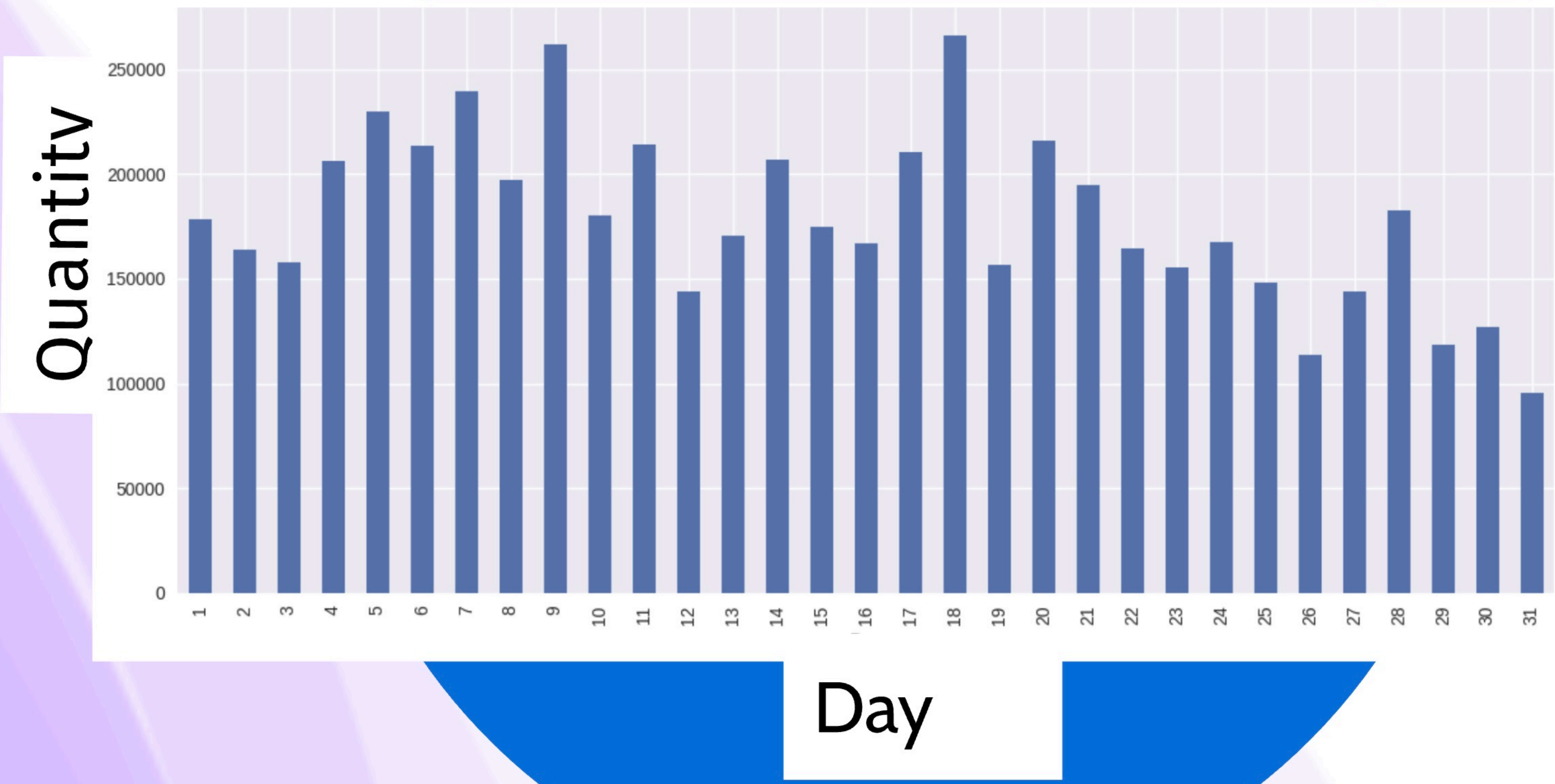
Order price

Sales

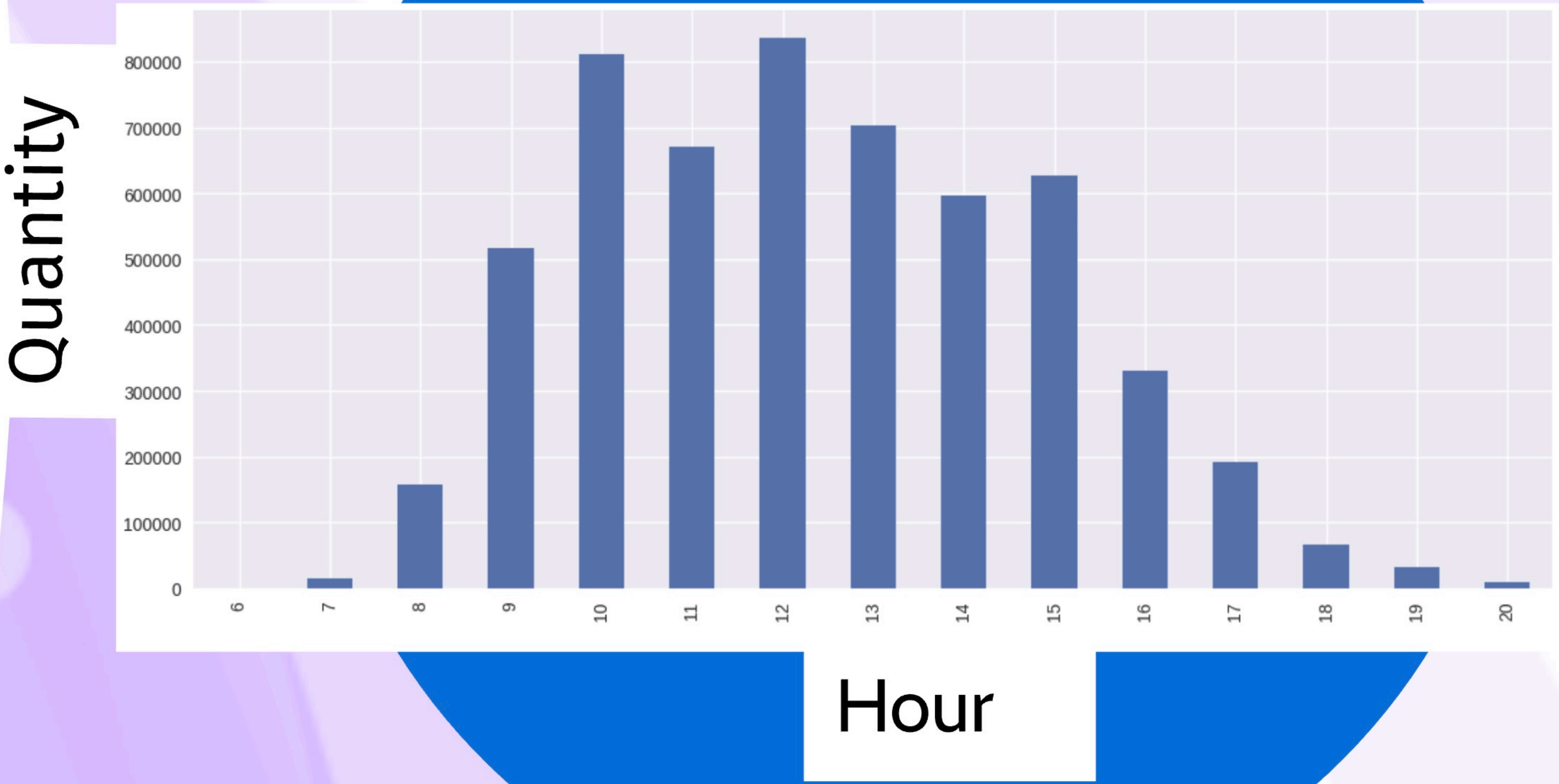

Quantity
700000
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
11
12

Month

Sales

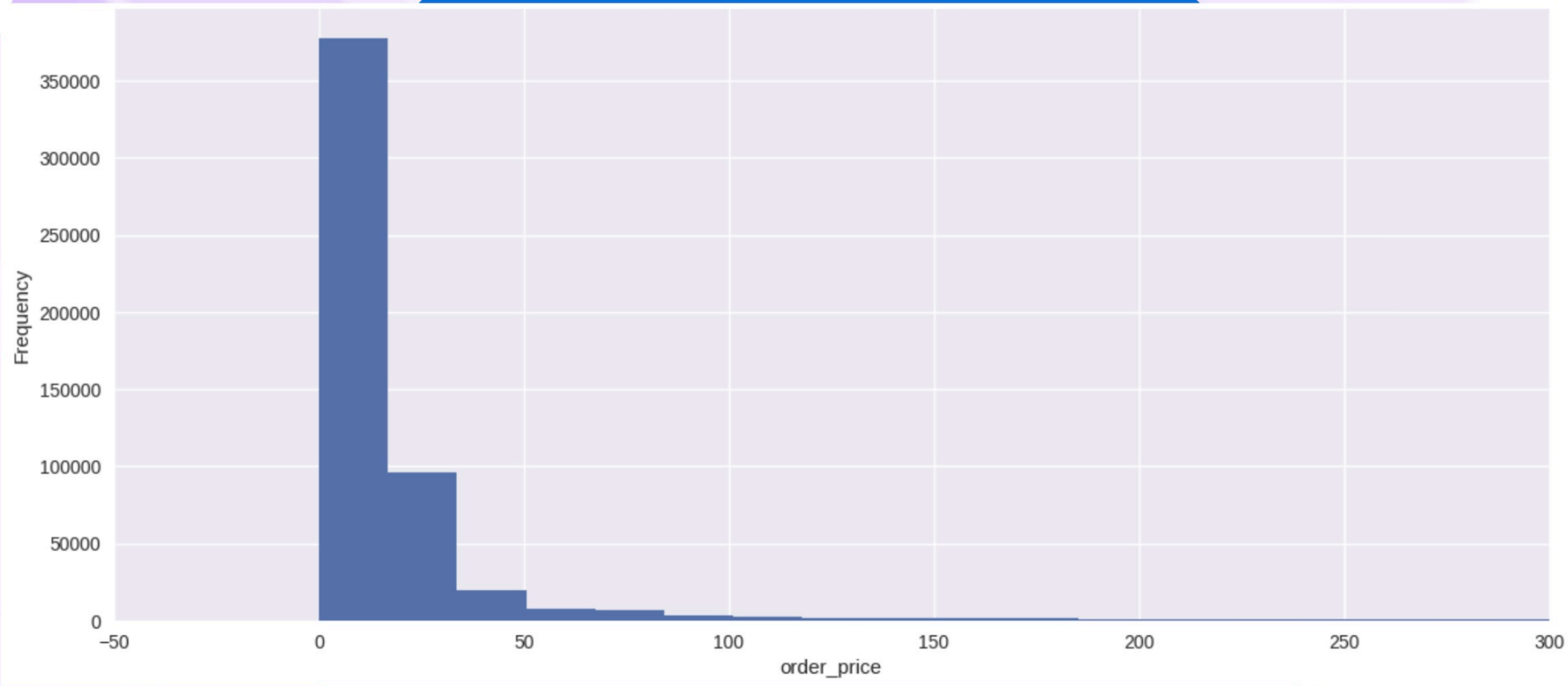

Quantity
250000
200000
150000
100000
50000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
Day

Sales



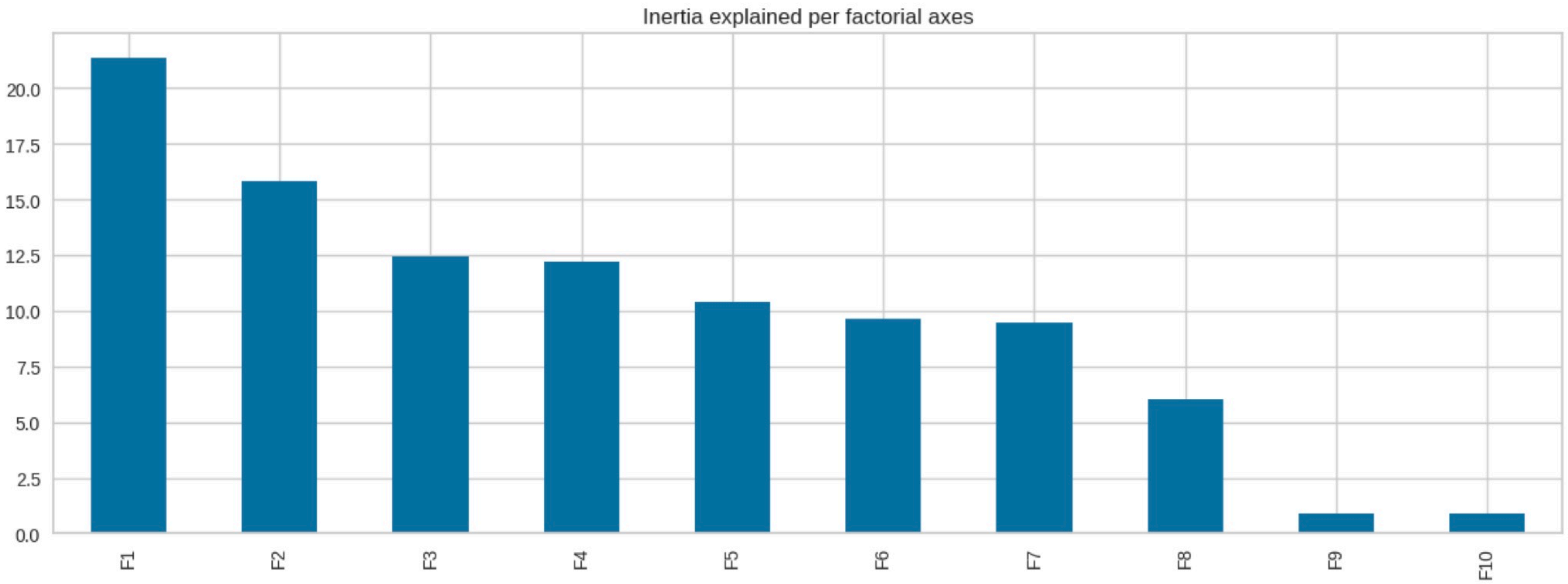
Quantity
800000
700000
600000
500000
400000
300000
200000
100000
0
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Hour

Order price



350000
300000
250000
200000
150000
100000
50000
0
Frequency
−50
0
50
100
150
200
250
300
order_price

PCA



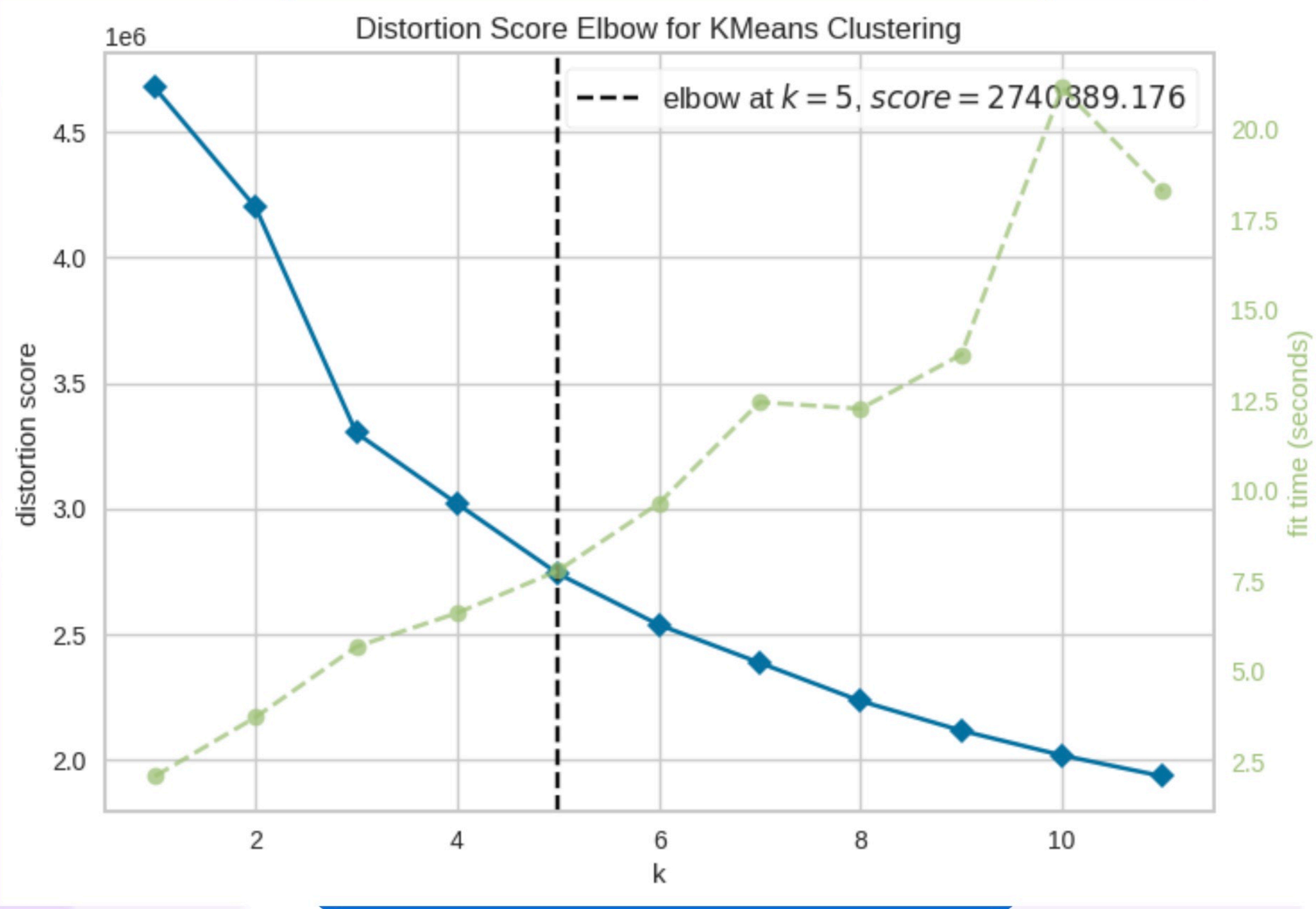
Choose of K

Interclass distance

First 8 eigen vector explain 97.4% of the variability

From 47 originals columns we keep only 8 columns

Elbow plot


Distortion Score Elbow for KMeans Clustering
1e6
elbow at $k = 5$, $score = 2740889.176$
distortion score
4.5
4.0
3.5
3.0
2.5
2.0
fit time (seconds)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
2
4
6
8
10
k