Exploratory Data Analysis on E-Commerce Data

To discover interesting transactional patterns of different customers and countries

Admond Lee Follow
Sep 23, 2018 · 9 min read

(Source)

In general explanation, data science is nothing more than using advanced statistical and machine learning techniques to solve various problems using data. Yet, it's easier to just

dive into applying some fancy machine learning algorithms —and Voila! You got the prediction — without first understanding the data.

This is exactly where the importance of **Exploratory Data Analysis (EDA)** (as defined by Jaideep Khare) comes in which, unfortunately, is a commonly undervalued step as part of the data science process.

EDA is so important for 3 reasons (at least) as stated below:

1. Make sure business stakeholders ask the right questions — often by exploring and visualizing data — and validate their business assumptions with thorough investigation
2. Spot any potential anomalies in data to avoid feeding wrong data to a machine learning model
3. Interpret the model output and test it's assumptions

There you have it. Now that we have already understood the **"WHAT and WHY"** aspects of EDA, let's examine a dataset together and go through the **"HOW"** that will eventually lead us to discover some interesting patterns, as we'll see in the next section.

We'll focus on the overall workflow of EDA, visualization and its results. For technical reference, please refer to **my notebook on Kaggle** anytime you want to have a more detailed understanding of the codes.

To give a brief overview, this post is dedicated to 5 sections as follow:

1. Context of Data
2. Data Cleaning (a.k.a data preprocessing)
3. Exploratory Data Analysis
4. Results
5. Conclusion

Let's get started and have fun!

• • •

Context of Data

(Source)

In this post, we'll investigate the E-Commerce dataset obtained from Kaggle. Before dealing with the dataset, let's try to understand what it is about to give us a better understanding of its context.

In short, the dataset consists of **transactional data with customers in different countries who make purchases from an online retail company based in the United Kingdom (UK) that sells unique all-occasion gifts**. The information is summarized as below:

- Company — UK-based and registered non-store online retail
- Products for selling — Mainly all-occasion gifts

- Customers — Most are wholesalers (local or international)
- Transactions Period — **1st Dec 2010–9th Dec 2011 (One year)**

. . .

Data Cleaning

(Source)

We all know data in real world is messy (including Kaggle!) and thus, let's spend some time to clean the data to the format we need. Below is a snapshot of what the original data looks like after loading the dataset into a dataframe.

	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	3.39	17850.0	United

As intuitive as the variables (column names) may sound, let's take a step further by understanding what each variable means:

InvoiceNo *(invoice_num)*: A number assigned to each transaction
StockCode *(stock_code)*: Product code
Description *(description)*: Product name
Quantity *(quantity)*: Number of products purchased for each transaction
InvoiceDate *(invoice_date)*: Timestamp for each transaction
UnitPrice *(unit_price)*: Product price per unit
CustomerID *(cust_id)*: Unique identifier each customer
Country *(country)*: Country name

NOTES → Product price per unit is assumed to follow the same currency throughout our analysis

```
# check missing values for each column
df.isnull().sum().sort_values(ascending=False)
```

```
cust_id         135080
description       1454
country              0
unit_price           0
invoice_date         0
quantity             0
stock_code           0
invoice_num          0
dtype: int64
```

Check missing values

So far, so good. We see that there are some missing values for Customers ID and Description. The rows with any of these missing values will therefore be removed.

```
df_new.describe().round(2)
```

	quantity	unit_price	cust_id

count	406829.00	406829.00	406829.00
mean	12.06	3.46	15287.69
std	248.69	69.32	1713.60
min	-80995.00	0.00	12346.00
25%	2.00	1.25	13953.00
50%	5.00	1.95	15152.00
75%	12.00	3.75	16791.00
max	80995.00	38970.00	18287.00

Descriptive statistic of data

By understanding the data in a more descriptive manner, we notice two things:

1. Quantity has negative values
2. Unit Price has zero values (FREE items?)

Interesting…

At this stage, we'll just remove Quantity with negative values — this notebook explains what negative values mean — and Unit Price with zero values will be explained in the later part.

To calculate the total money spent on each purchase, we simply multiply Quantity with Unit Price:

$$amount\_spent = quantity * unit\_price$$

Finally, we add a few columns that consist of the Year_Month, Month, Day and Hour for each transaction for analysis later. The final dataframe will look like this:

	invoice_num	invoice_date	year_month	month	day	hour	stock_code	description	quantity	unit_price	amount_spent	cust_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

Final dataframe

. . .

Exploratory Data Analysis

(Source)

Highest number of orders and money spent on purchases

	cust_id	country	invoice_num
4019	17841	United Kingdom	7847
1888	14911	EIRE	5677
1298	14096	United Kingdom	5111
334	12748	United Kingdom	4596
1670	14606	United Kingdom	2700

Top 5 customers with most number of orders

	cust_id	country	amount_spent

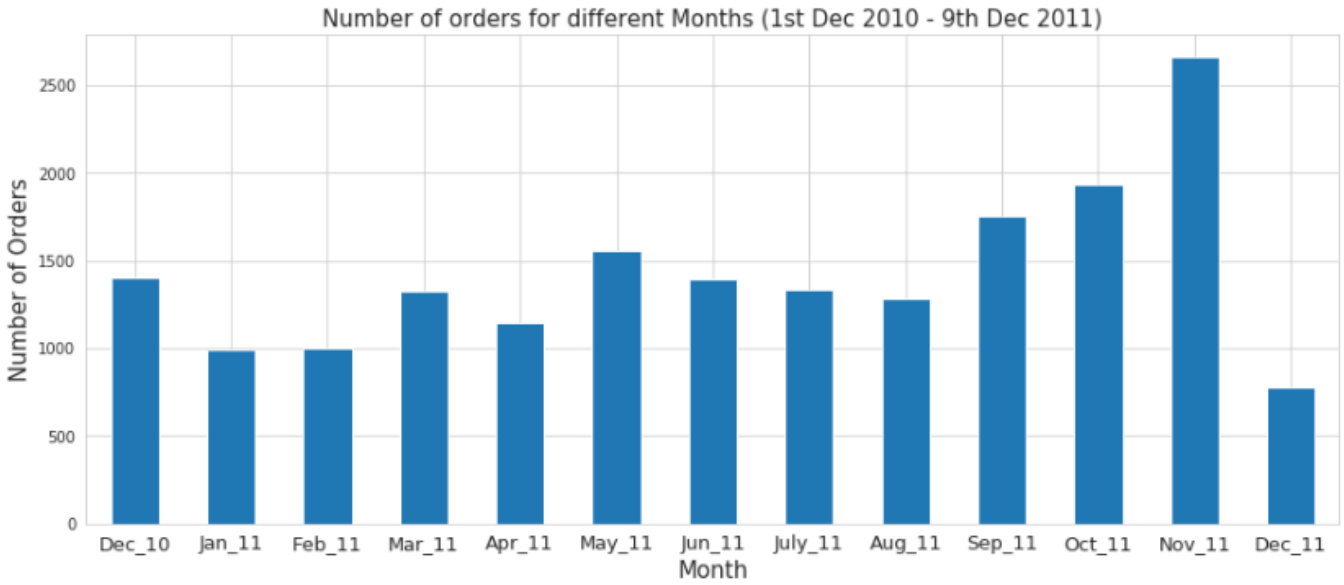
1698	14646	Netherlands	280206.02
4210	18102	United Kingdom	259657.30
3737	17450	United Kingdom	194550.79
3017	16446	United Kingdom	168472.50
1888	14911	EIRE	143825.06

Top 5 customers with highest money spent

In E-Commerce world, we often want to know which customers — where they come from—place the most orders and spend the most money as they drive the sales of companies.

From the results we observe that most orders are made in the UK and customers from Netherlands spend the highest amount of money in their purchases.

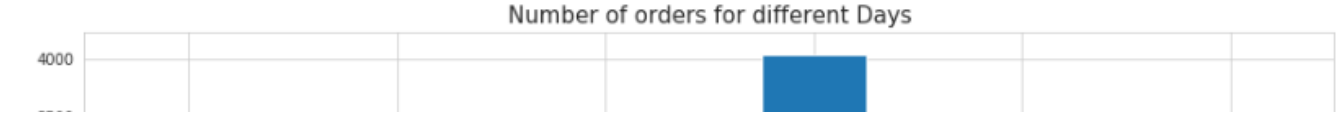
How many orders (per month)?

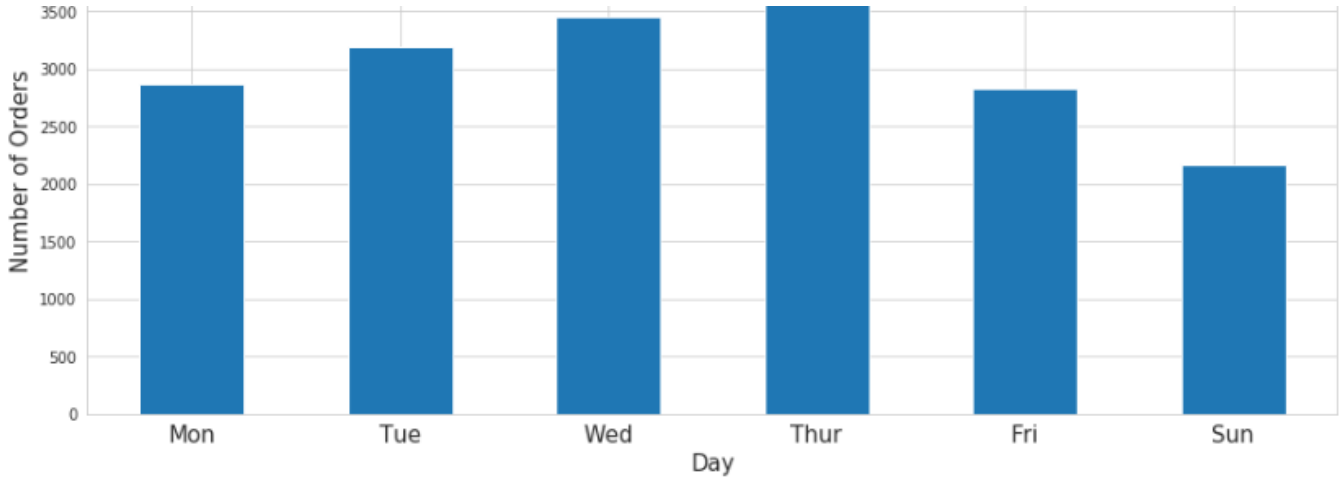


Number of orders for different months

Overall, we consider that the company receives the highest number of orders in November 2011 since we do not have the full month of data for December 2011.

How many orders (per day)?



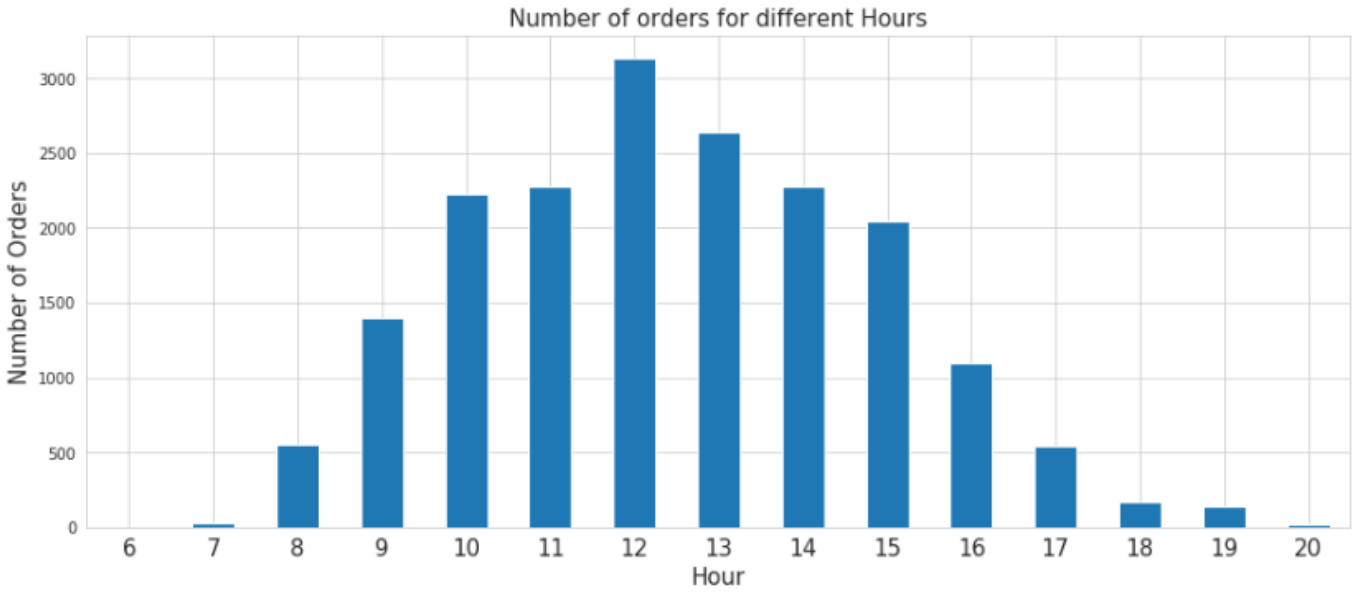


Number of orders for different days

Surprisingly, there are no transactions on Saturday throughout the whole period (1st Dec 2010–9th Dec 2011). Reasons behind are left for discussion as the dataset and its context are limited.

We also spot a trend where the number of orders received by the company tends to increases from Monday to Thursday and decrease afterward.

How many orders (per hour)?



Number of orders for different hours

In terms of hours, there are no transactions after 8:00pm until the next day at 6:00am.

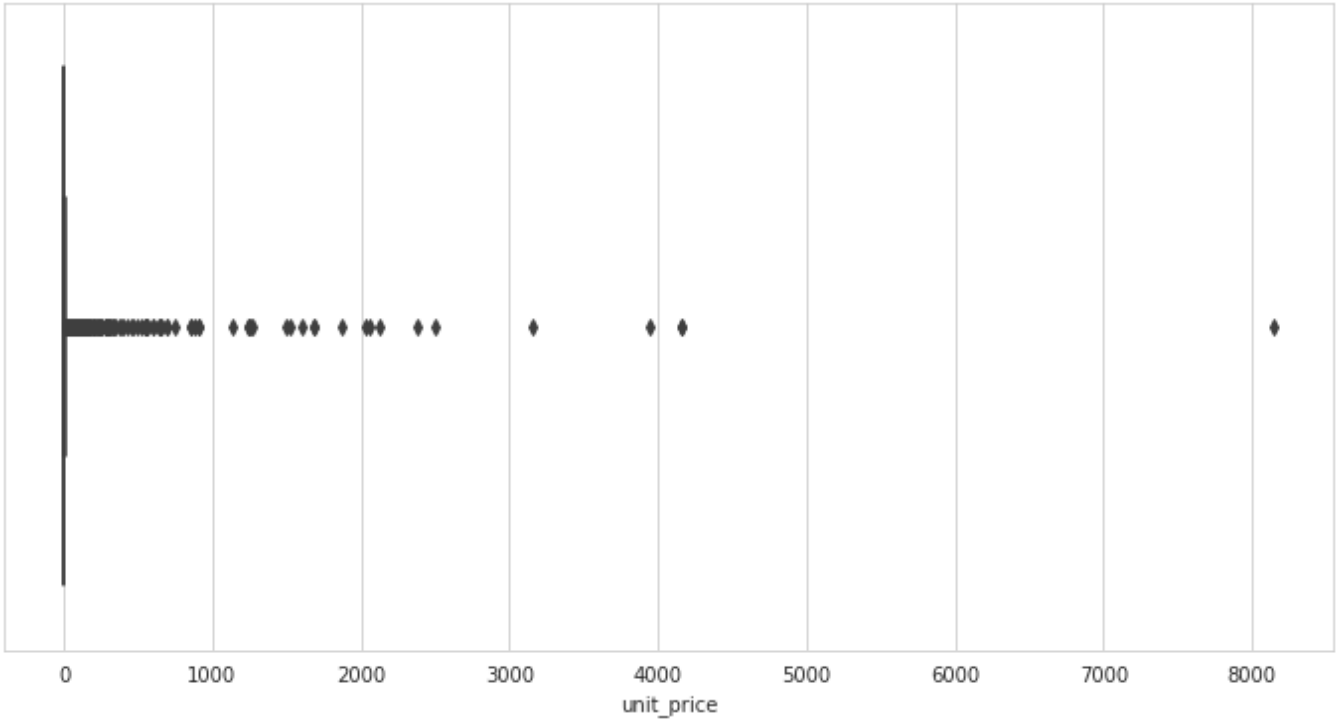
Besides, we notice that the company receives the highest number of orders at 12:00pm. One of the reasons could be due to the fact that most customers make purchases during lunch hour between 12:00pm — 2:00pm.

Discover transactional patterns for Unit Price

```
df_new.unit_price.describe()
```

```
count    397924.000000
mean          3.116174
std          22.096788
min           0.000000
25%           1.250000
50%           1.950000
75%           3.750000
max        8142.750000
Name: unit_price, dtype: float64
```

Descriptive statistics of Unit Price

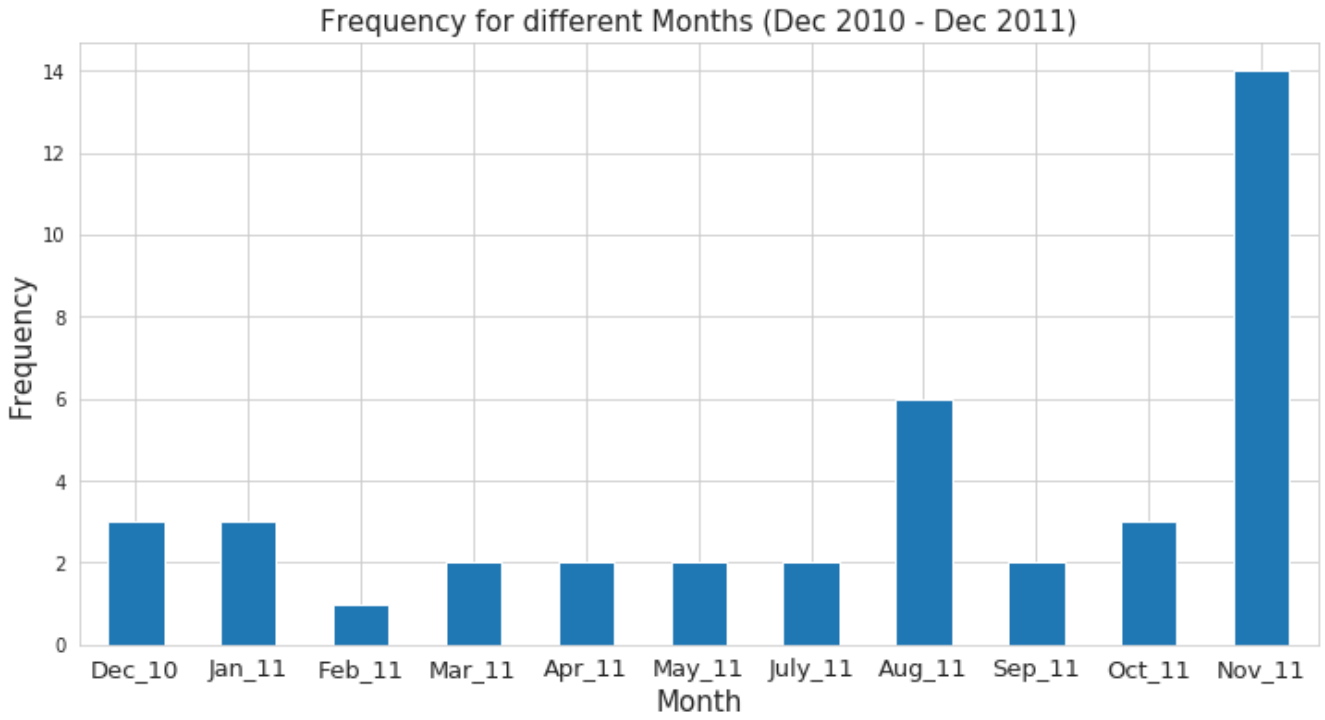


Boxplot for Unit Price

Before we move our attention to the zero values (FREE items) of unit price, we make a boxplot to check the distribution of the unit price for all products.

We observe that 75% of the data has unit price of less than 3.75 dollars — which indicates most products are relatively cheap. Only minority of them has high prices per unit (Again, we assume each price per unit follows the same currency).

Well... FREE items for purchase? YES, maybe...



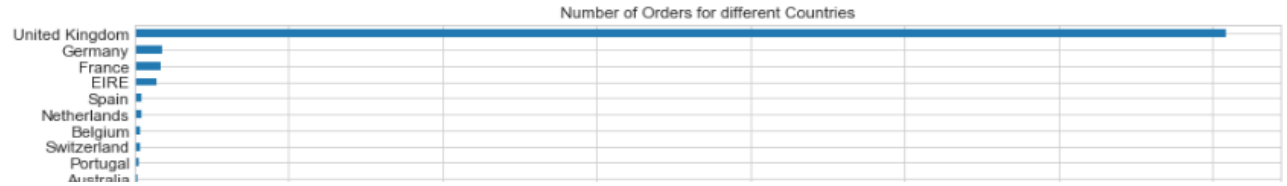
Frequency of giving out FREE items for different months

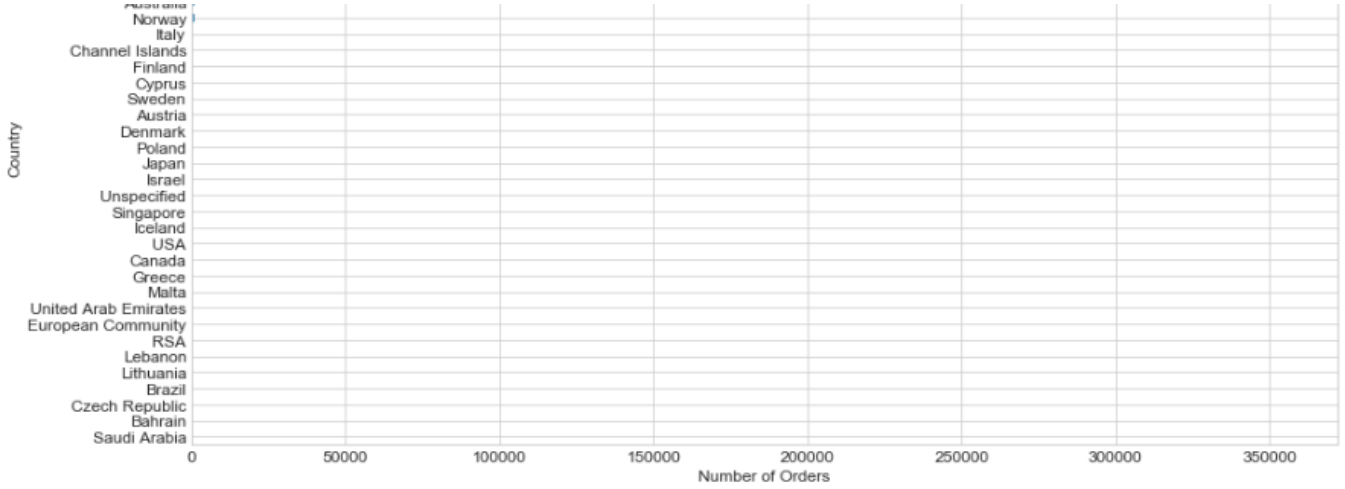
From the plot, the company tends to **give out FREE items for purchases occasionally each month (except June 2011)**.

However, it is not clear what factors contribute to giving out the FREE items to the particular customers. More in-depth analysis could be done for further explanation. Let me know if you have found out the reasons behind! 😉

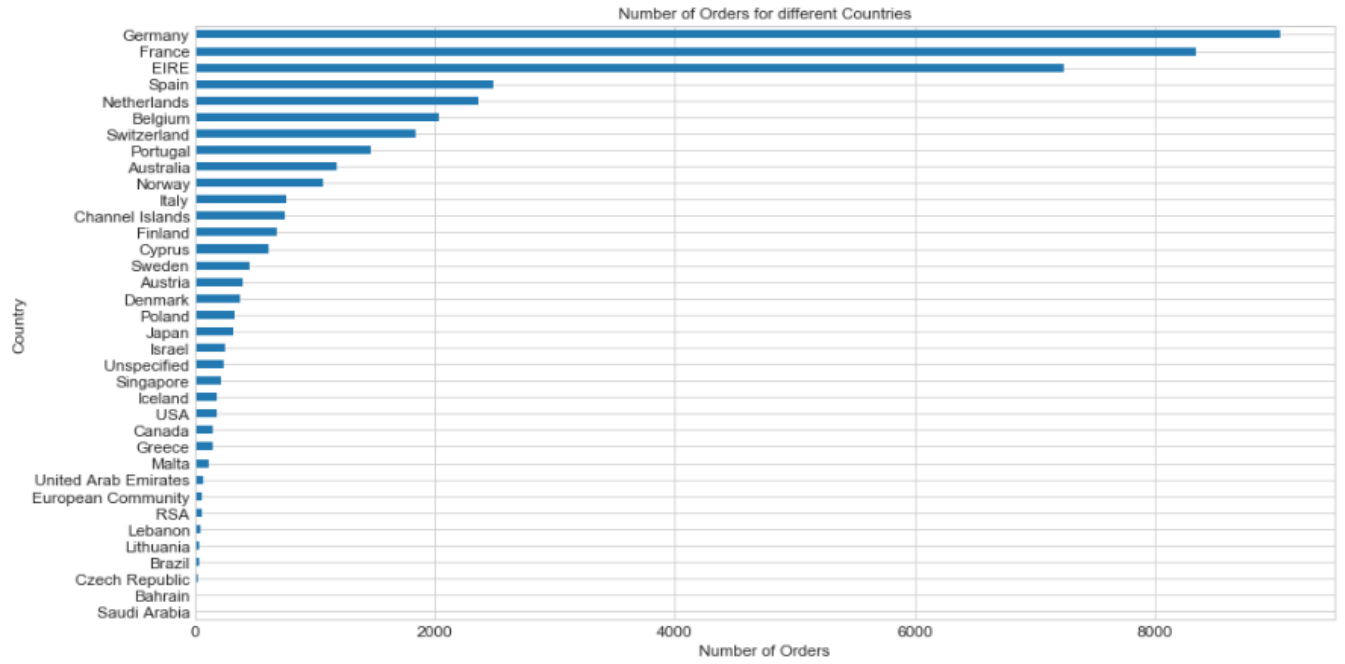
Discover transactional patterns for each Country

Top 5 countries with most number of orders





Number of orders in each country **(with UK)**



Number of orders in each country **(without UK)**

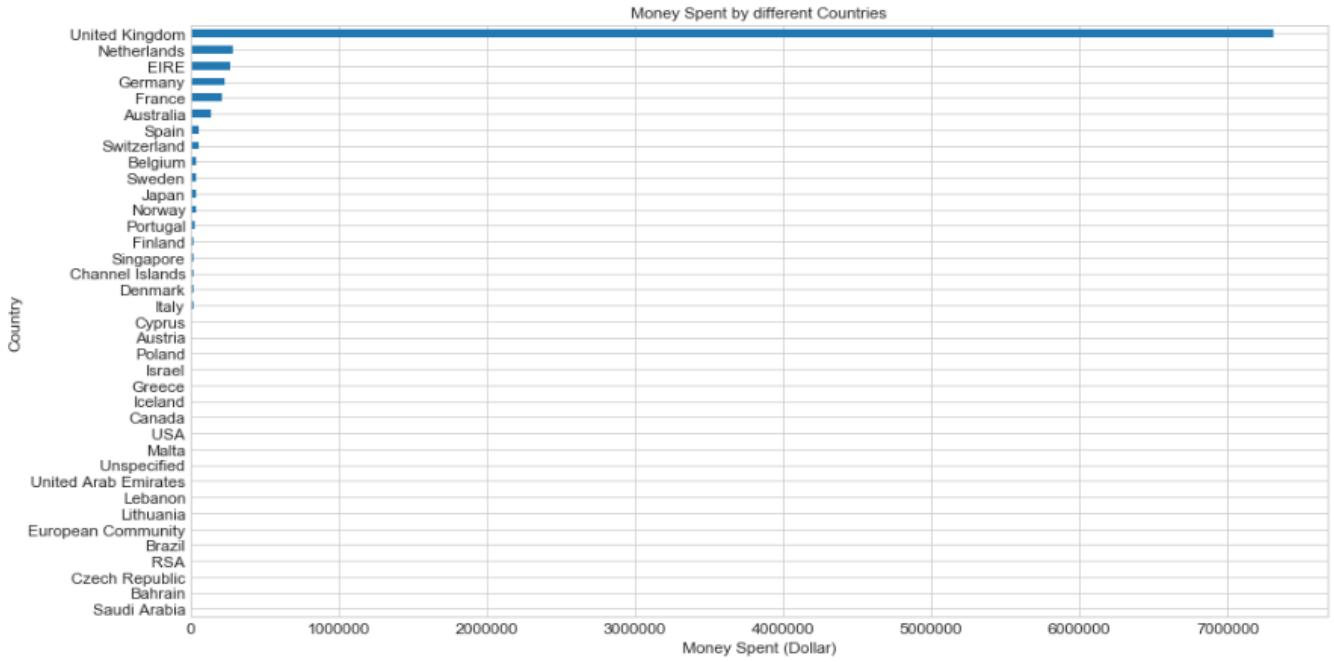
As expected, the company receives the highest number of orders in the UK (since it is a UK based company).

To better discern the trend, UK is removed for clearer comparison among other countries. As a result, the TOP 5 countries (including UK) that place the highest number of orders are as below:

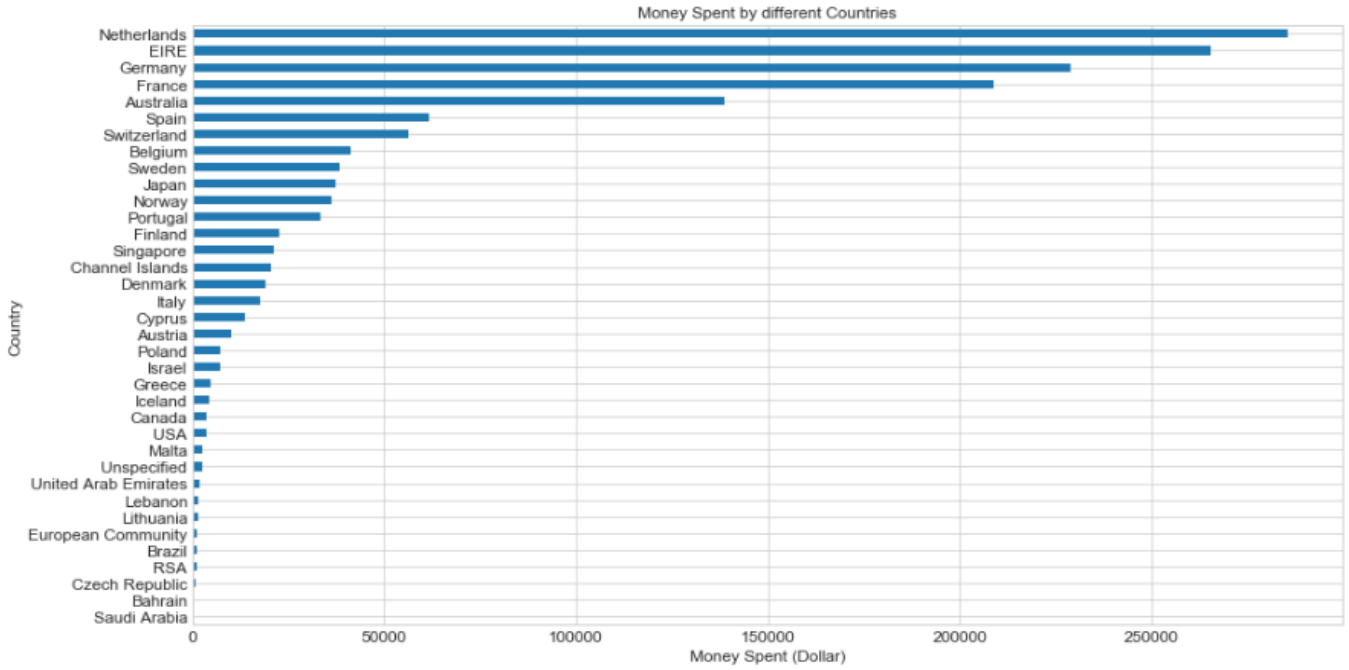
- United Kingdom
- Germany

- France
- Ireland (EIRE)
- Spain

Top 5 countries with highest money spent



Total money spent by each country **(with UK)**



Total money spent by each country **(without UK)**

As the company receives the highest number of orders from customers in the UK, it is natural to see that customers in the UK spend the most on their purchases.

Same as before, UK is removed for clearer comparison among other countries. The TOP 5 countries (including UK) that spend the most money on purchases are as below:

- United Kingdom
- Netherlands
- Ireland (EIRE)
- Germany
- France

. . .

Results from EDA

1. The **customer with the highest number of orders comes from the United Kingdom (UK)**
2. The **customer with the highest money spent on purchases comes from Netherlands**
3. The company receives the highest number of orders from customers in the UK (since it is a UK-based company). Therefore, the **TOP 5 countries** (including UK) that place the **highest number of orders** are as follow → **United Kingdom, Germany, France, Ireland (EIRE), Spain**
4. As the company receives the highest number of orders from customers in the UK (since it is a UK-based company), customers in the UK spend the most on their purchases. Therefore, the **TOP 5 countries** (including UK) that **spend the most money on purchases** are as follow → **United Kingdom, Netherlands, Ireland (EIRE), Germany, France**
5. **November 2011 has the highest sales.** The month with the lowest sales is undetermined as the dataset consists of transactions until 9th December 2011 in

December

6. There are **no transactions on Saturday** between 1st Dec 2010 — 9th Dec 2011
7. The number of orders received by the company tends to increases from Monday to Thursday and decrease afterward
8. The company receives the **highest number of orders at 12:00pm.** Possibly most customers made purchases during **lunch hour between 12:00pm — 2:00pm**
9. The company tends to **give out FREE items for purchases occasionally each month (Except June 2011).** However, it is not clear what factors contribute to giving out the FREE items to the particular customers

. . .

Conclusion

(Source)

Awesome!

Simply by performing EDA on the dataset we've identified some interesting results. Of course, the results don't just stop here. They can always be used to validate business assumptions (if any) and interpret a machine learning model's output and so much more!

Remember. **Creativity is your limit when doing EDA**. And it really depends on your business understanding, curiosity to ask interesting questions to challenge and validate assumptions, as well as your intuition.

Thank you for reading. Hopefully by showing the overall workflow of EDA, visualization and its results, EDA will become less intimidating to you and you'll be more interested in getting your hands dirty next time.

As always, if you have any questions or comments feel free to leave your feedback below or you can always reach me on LinkedIn. Till then, see you in the next post! 😄

About the Author

Admond Lee is now in the mission of making data science accessible to everyone. He is helping companies and digital marketing agencies achieve marketing ROI with actionable insights through innovative data-driven approach.

With his expertise in advanced social analytics and machine learning, Admond aims to bridge the gaps between digital marketing and data science.

Check out his **website** if you want to understand more about Admond's story, data science services, and how he can help you in marketing space.

You can connect with him on LinkedIn, Medium, Twitter, and Facebook.

Admond Lee

In the mission of making data science accessible to everyone. Admond is helping companies and digital marketing agencies achieve their marketing R...

www.admondlee.com

Data Science Towards Data Science Exploratory Data Analysis Education Data Visualization

About Help Legal