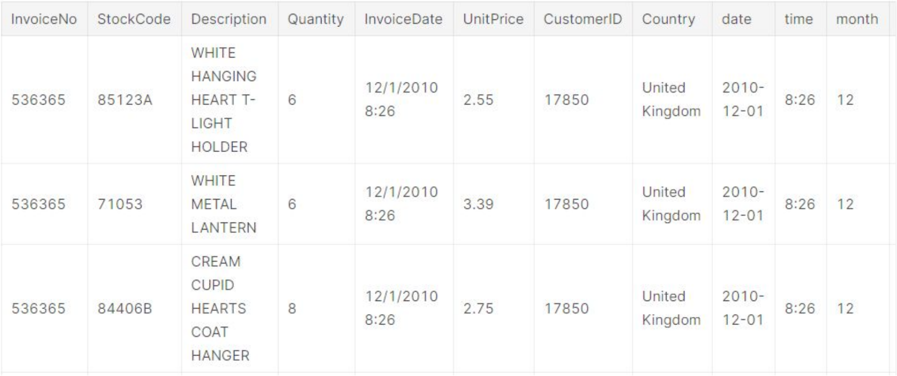
E-commerce transaction data analysis and Customer segmentation

In this project, I will be analysing customer purchase and product sales history, I can group products and customers into groups that behave similarly, and make data-driven business decisions that can improve a wide range of inventory and sales key performance indicators (KPIs).

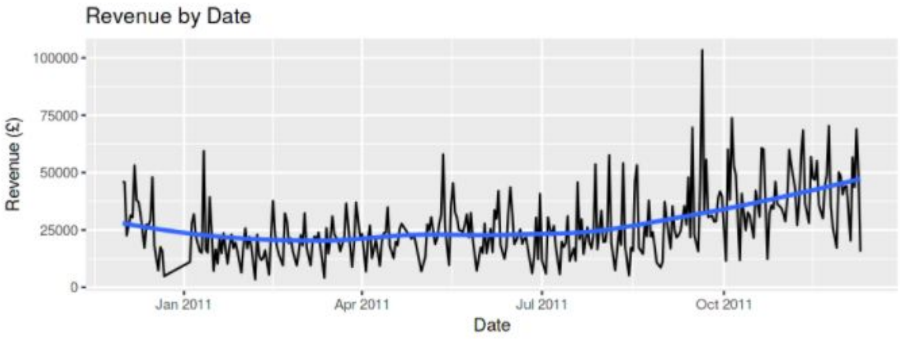
I begin with cleaning the data, for that first I will drop all the missing value rows, since its a good-sized set of data, dropping rows all together shouldn't be a problem.

Next is creating two new variables: date and time from InvoiceDate variable. Also create a new column by multiplying the Quantity by the UnitPrice for each row. Our last step would be turn the appropriate variables into factors
This is what my final Dataframe looks like:



InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	date	time	month
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom	2010-12-01	8:26	12
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom	2010-12-01	8:26	12
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom	2010-12-01	8:26	12

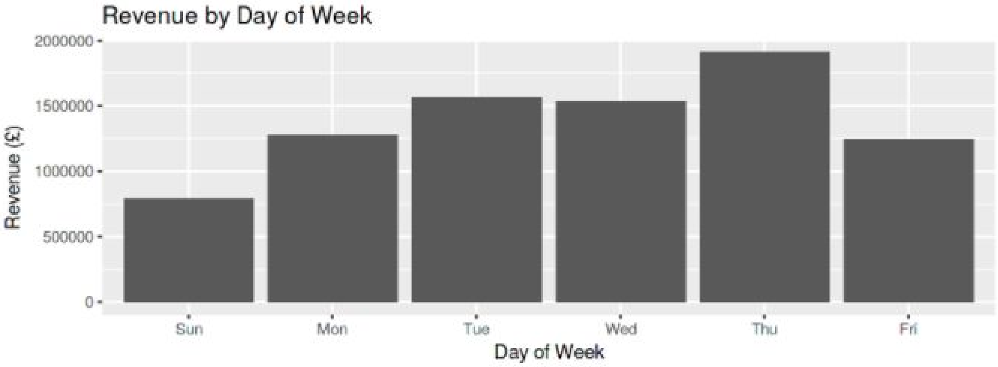
Let’s begin with the Analysis:



From the above plot we can see that sales are trending up, which is good. Let’s look for more information.

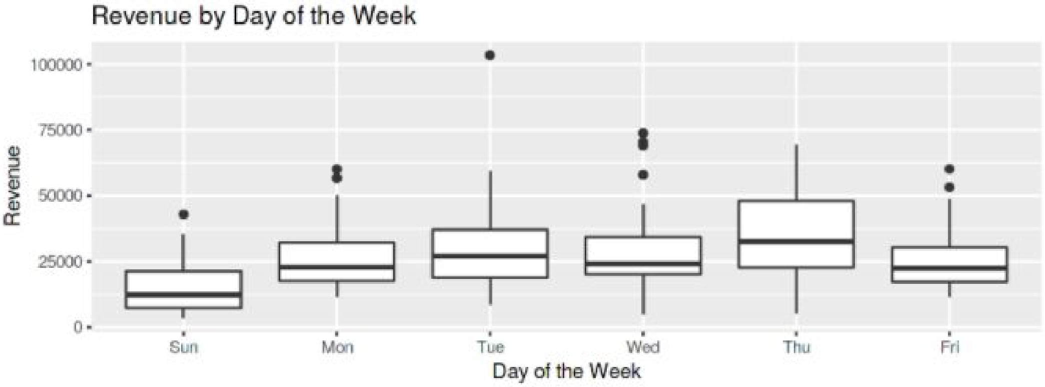
Day of week analysis

Next goal is to find out if the days of the week side of our data can give us any interesting insight.



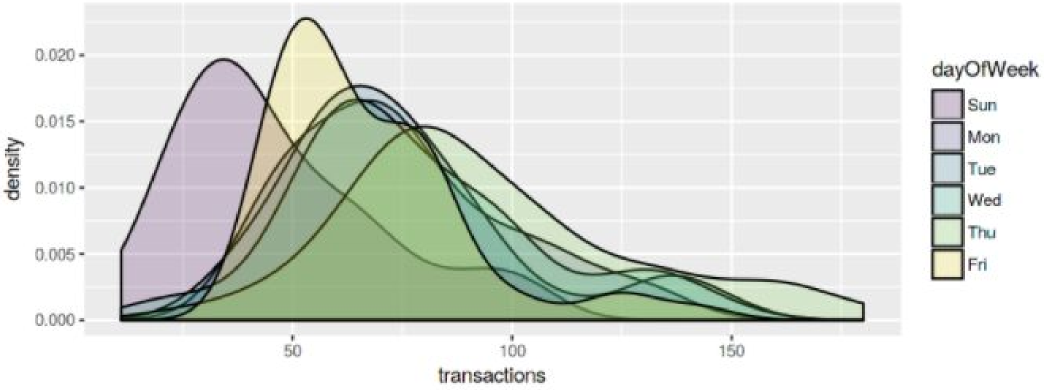
It looks like there could be something interesting going on with the amount of revenue that is generated on each particular weekday. Let's drill into this a little bit more by creating a new dataframe that I can use to look at what's going on at the day of the week level in a bit more detail

After doing some grouping, the dataset is ready to do some more detailed analysis on each day



The plot shows some difference in the amount of revenue generated on each day of the week.

There appears to be a strong correlation between the number of transactions and the revenue generated, the fluctuations in both the graphs are almost the same.



There appears to be a reasonable amount of skewness in our distributions. We can use a non-parametric test to look for statistically significant differences in our data.

Conclusions from our day-of-the-week summary

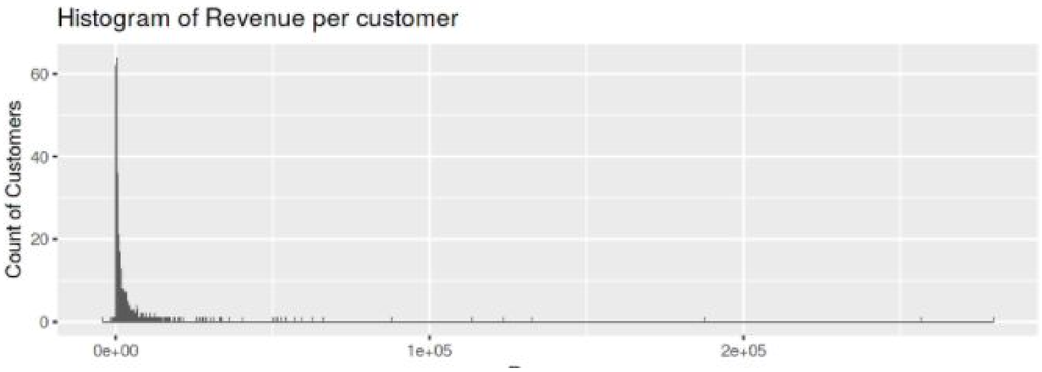
By looking at our data at the level of weekdays, I can see that there are statistically significant differences in the number of transactions that take place on different days of the week, with Sunday having the lowest number of transactions, and Thursday the highest.

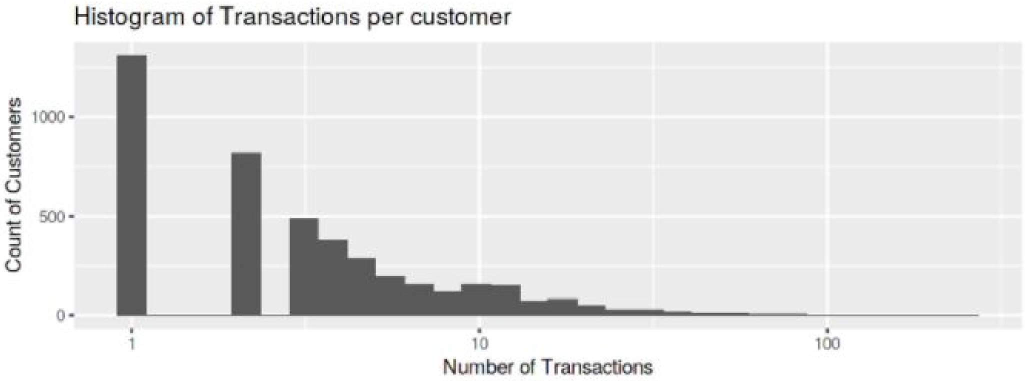
Given the low number of transactions on a Sunday and a high number on a Thursday, I could make recommendations around our advertising spend. Should I spend less on a Sunday and more on a Thursday, given that I know I already have more transactions, which could suggest people are more ready to buy on Thursdays? Possible, but without knowing other key metrics, it might be a bit hasty to say.

Well let's try to get some more information. In particular, some questions which i can think of right now are, How do these data correlate with web traffic figures? Does the conversion rate change or is there just more traffic on a Thursday and less on a Sunday?What about our current advertising spend? Is the company already spending less on a Sunday and more on a Thursday and that is behind our observed differences? What about buying cycles? How long does it take for a customer to go from thinking about buying something to buying it?

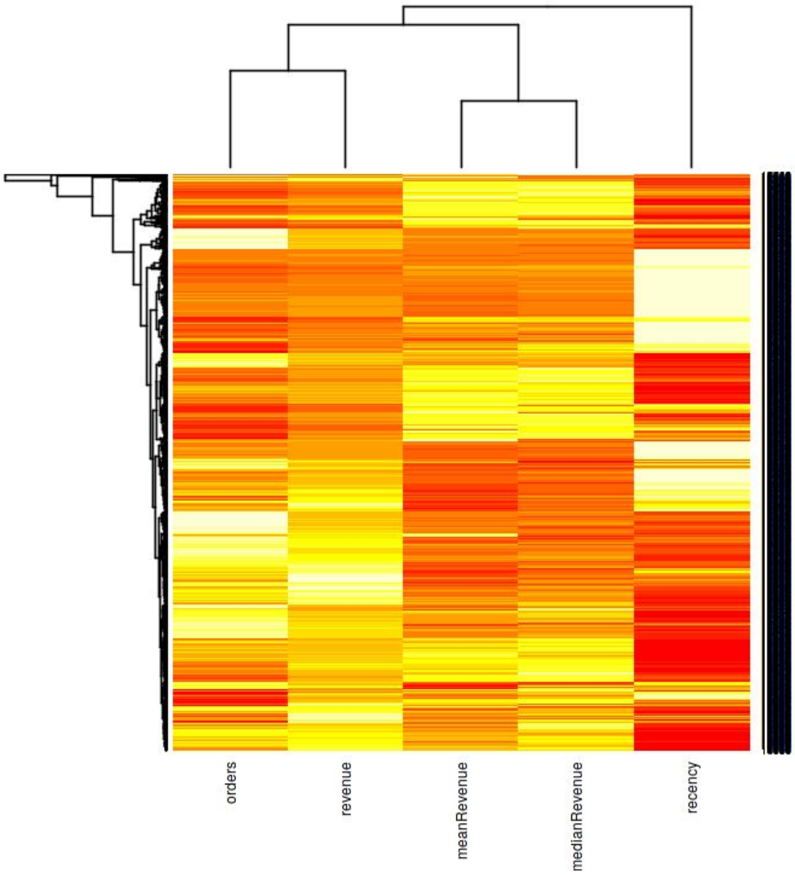
Customer segmentation

Last thing I want to do is divide my customers into clusters , see if we can find some similarity. This could help us more effectively target costumes via various customized deals and ads to reach our goal of higher revenue.





From the above plot we can infer that a small set of customer are the frequent customers , perhaps this small group is generating the maximum revenue.



In the heatmap, I can see that the total revenue clusters with the number of orders as I would expect, that the mean and median order values cluster together, again, as expected, and that the order recency sits in its own group. However, the main point of interest here is how the rows (customers) cluster.

By analysing the data in this way, I can uncover groups of customers that behave in similar ways. This level of customer segmentation is useful in marketing to these groups of customers appropriately. A marketing campaign that works for a group of customers that places low value orders frequently may not be appropriate for customers who place sporadic, high value orders for example.