E-commerce Dataset Exploration

Using R & Tableau

Purpose of exploration: To extract any actionable and relevant data from data sample and provide findings and recommendations.

About the data: This dataset appears to be from an Orders database table. It contained 541909 records for 8 fields -

- InvoiceNo: *Order number*
- StockCode: *Item stock code*
- Description: *Item description*
- Quantity: *Quantity of item*
- InvoiceDate: *Date of order - with Day/Month written in USA format*
- UnitPrice: *Price of item*
- CustomerID: *Customer Identifier*
- Country: *Order country*

Added:

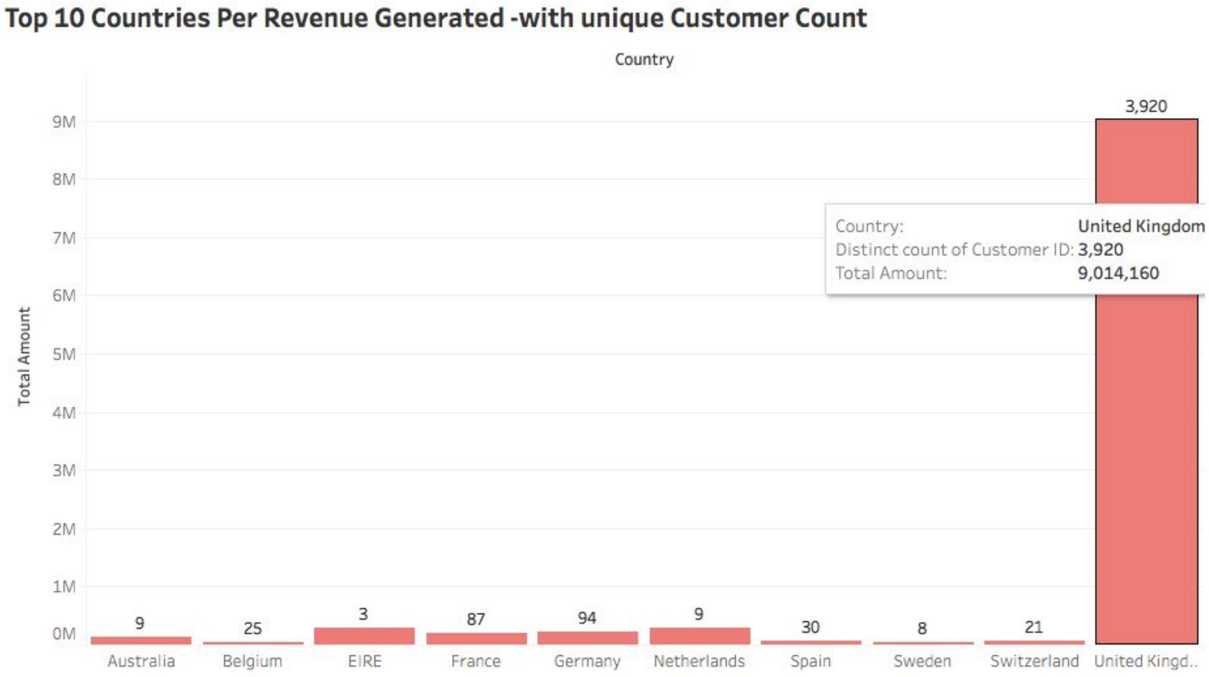
- Total Amount: *Quantity of item * Unit Price*
- Month: *Month of order*
- Day of the Week: *Day of order*
- Year: *Year of order*
- Time of the Day: *Hour of day of order*

This document covers some of the Findings (1A-F), a Summary (2) and Data Pre-processing steps (3).

1. Findings

A. *Most Revenue Generating Countries - Top 10*

From the 34 unique Country values in the dataset - this diagram displays the Top 10 per revenue. The United Kingdom is by far the most revenue generating country.



***NOTE:** The number on top of each block represents the number of unique customer in which that revenue was generated.*

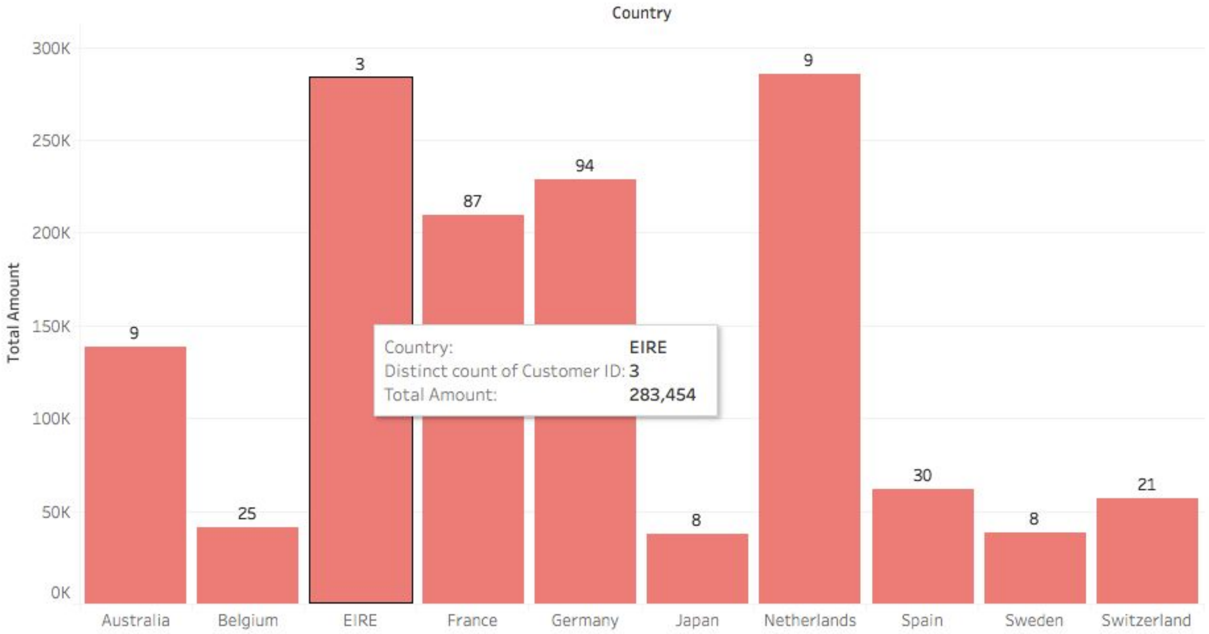
For the UK - this means an average of approx $2,299 per customer.

Now excluding the UK - taking a closer look at the Top 10, in the diagram below.

We can see that Eire(Ireland) and the Netherlands are the next two most revenue generating - this is especially interesting as they have a very low amount of customers for this revenue. Ireland has just 3 and the Netherlands has 9.

I noted that Eire has approx. 100 CustomerIDs missing.

Top 10 Countries Per Revenue Generated -with unique Customer Count

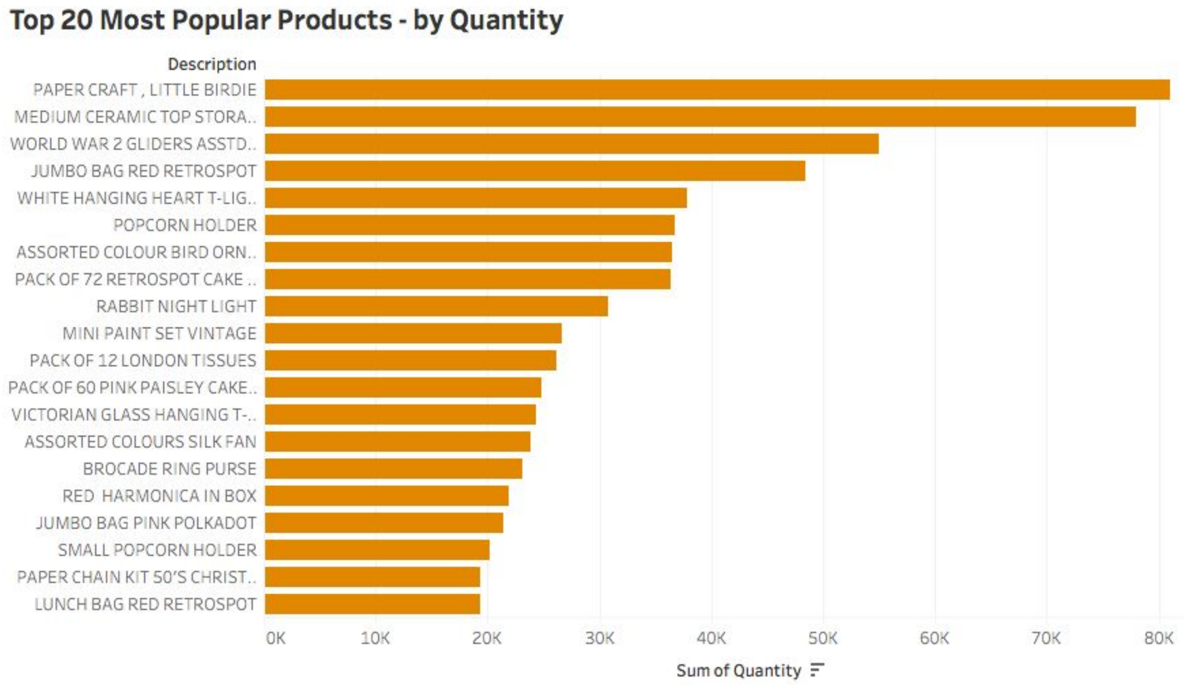


Based on these results - a small number of customers can make up a large amount of Revenue. Customer Retention for these countries could be a focus - and also looking for new customers who plan large orders.

B. Most Popular Products - Top 20

Purely looking at which Products have been sold the most by Quantity - the diagram below shows:

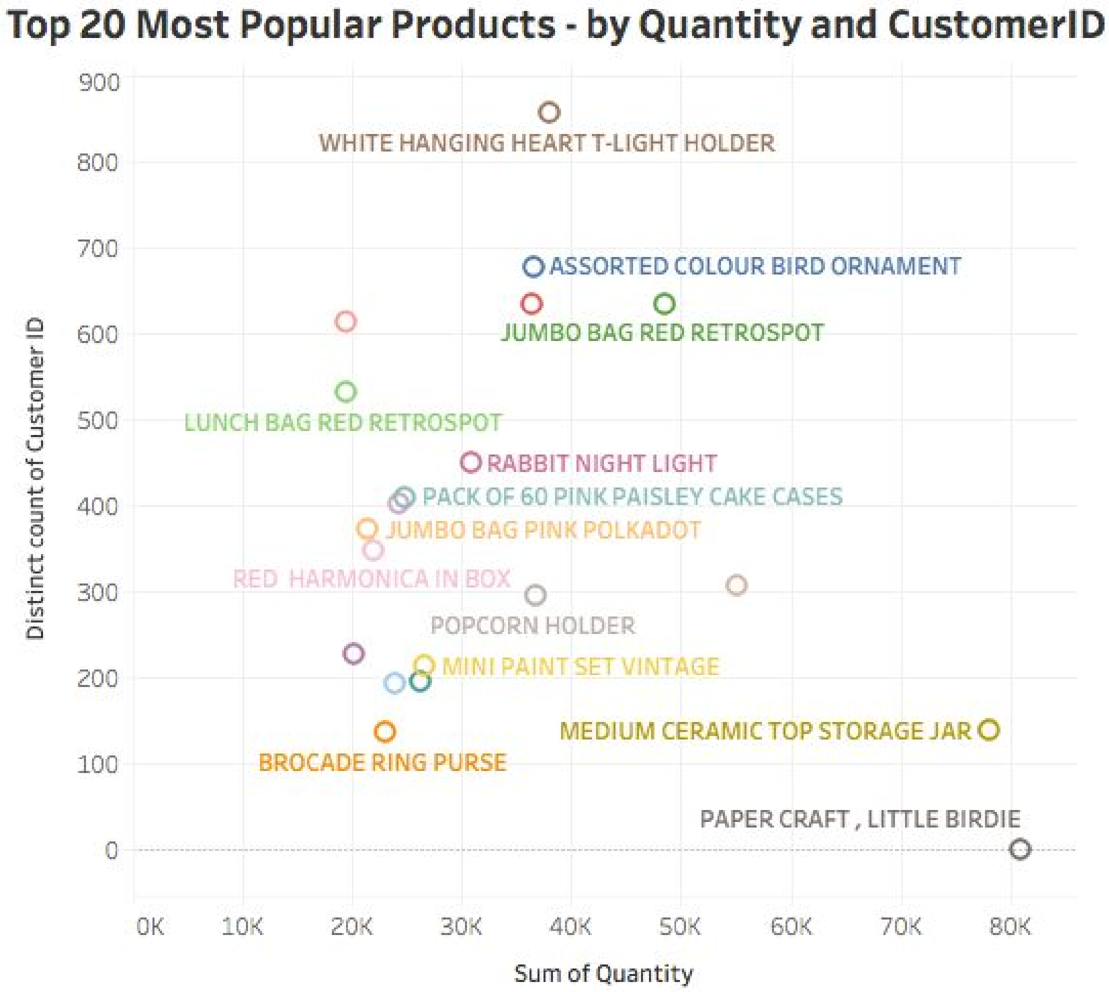
‘PAPER CRAFT, LITTLE BIRDIE’ and ‘MEDIUM CERAMIC TOP STORAGE JAR’ are the most Popular Products - way ahead of the other Top 18.



When another dimension is added, unique CustomerID, the data is represented quite differently. It appears that ‘PAPER CRAFT, LITTLE BIRDIE’ is a complete outlier - as there is only 1 unique CustomerID. ‘MEDIUM CERAMIC TOP STORAGE JAR’ also has a low customer base.

The most popular items per unique customers orders are:

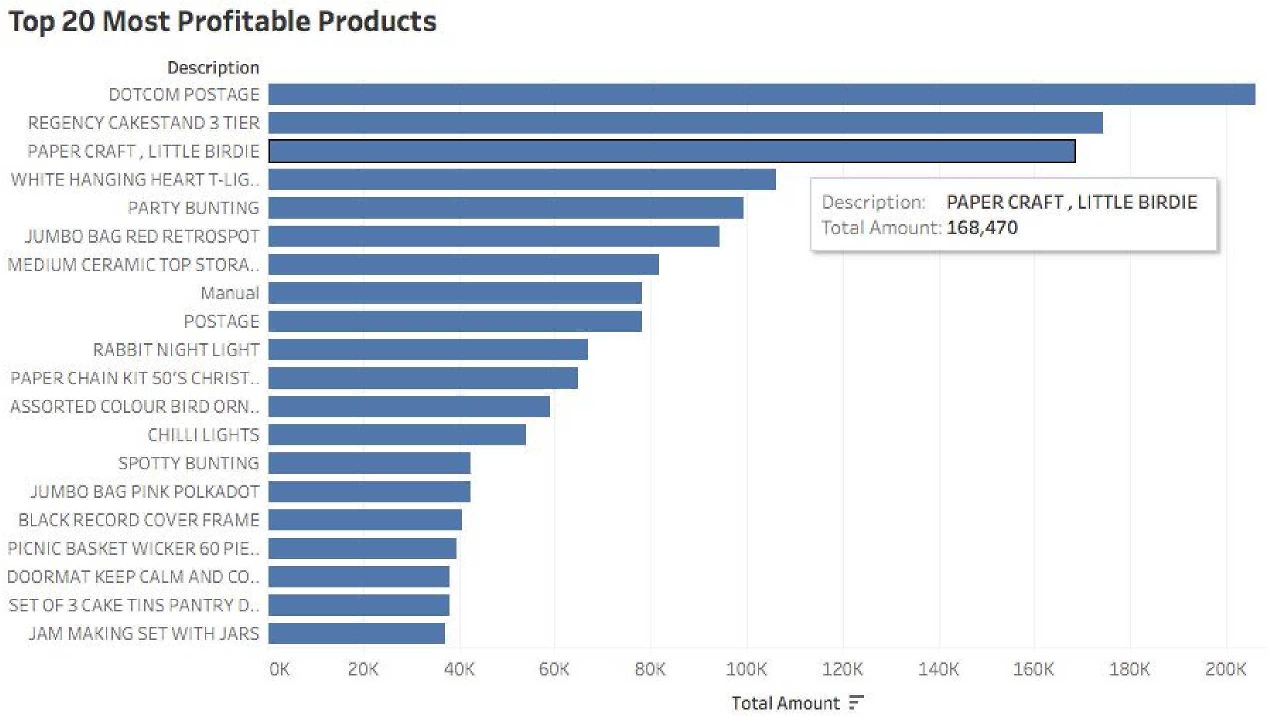
- WHITE HANGING HEART T-LIGHT HOLDER
- ASSORTED COLOUR BIRD ORNAMENT
- JUMBO BAG RED RETROSPOT



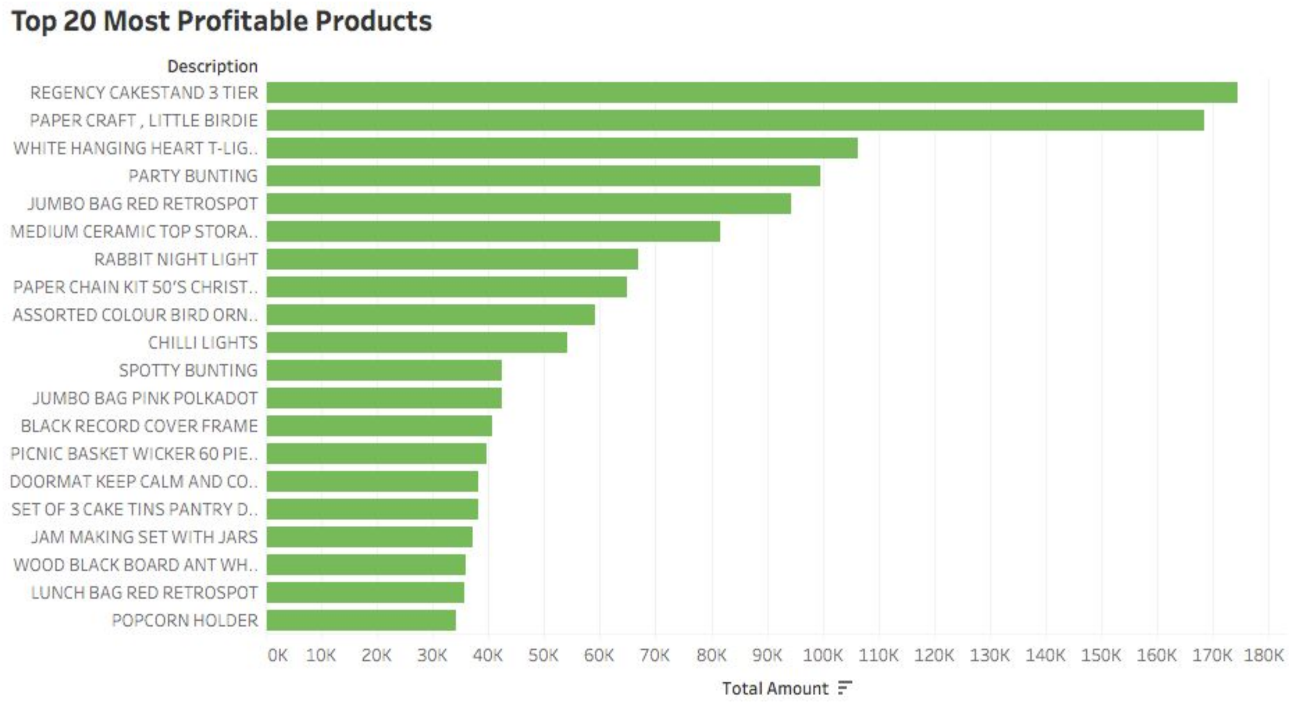
Top 20 Most Popular Products - by Quantity and CustomerID

Knowing your most popular products and the average quantity of sales per week/month can allow you to predict necessary stock levels.

C. Most Profitable Products - Top 20



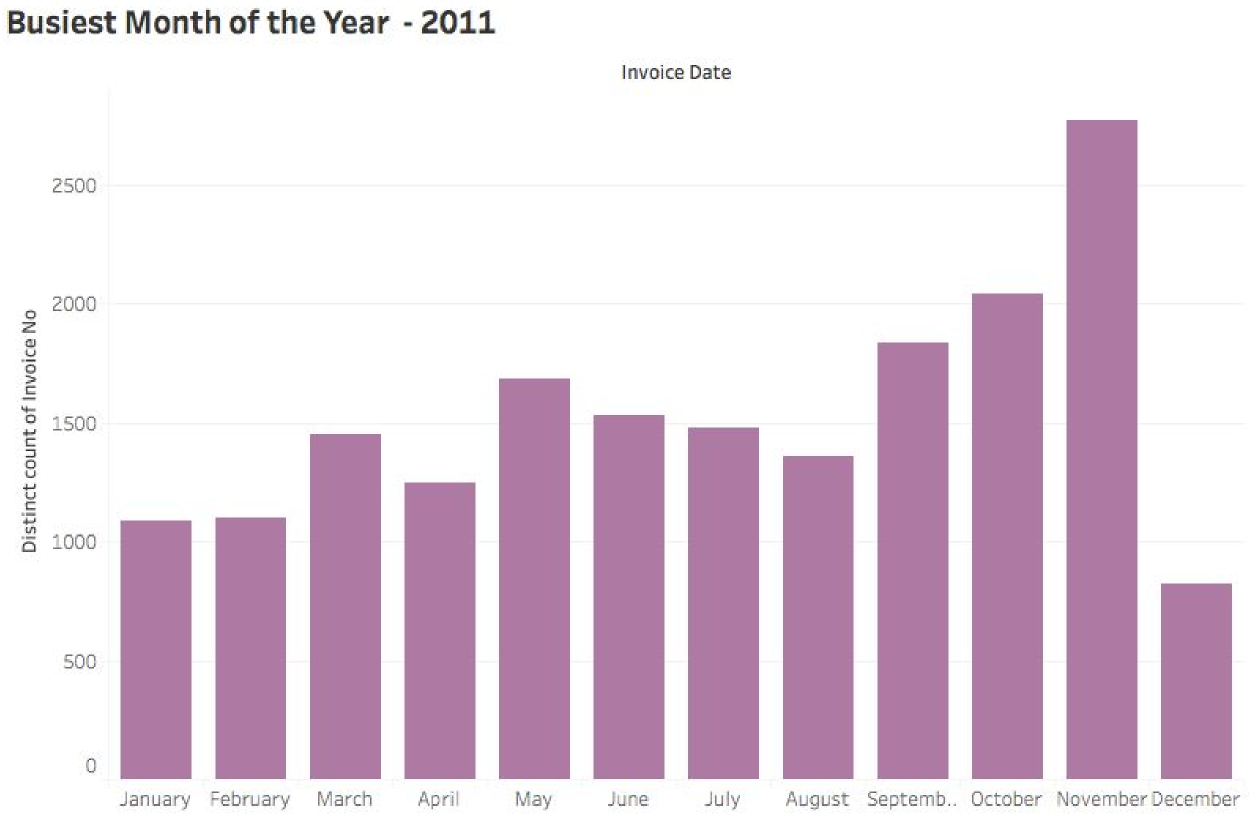
In the diagram above - ‘DOTCOM POSTAGE’, ‘Manual’ and ‘POSTAGE’ do not sound like actual products that the business sells. Therefore I excluded them for the diagram below. This can be easily adjusted once these products are confirmed as genuine.



‘REGENCY CAKESTAND 3 TIER’ is the most profitable Product.

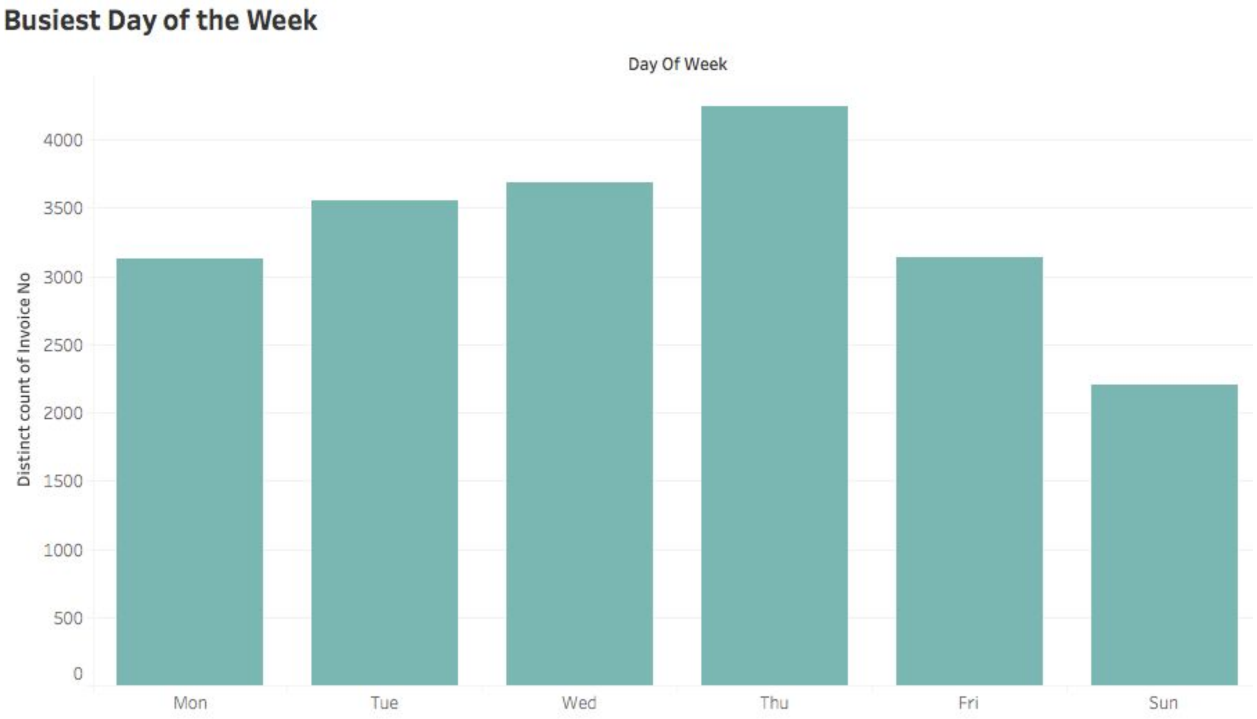
D. Busiest Time of the Year - Seasonality check

September, October and November show an upward trend in quantity of Orders. As these Products may be gifts - this makes sense with gift giving season being in December.



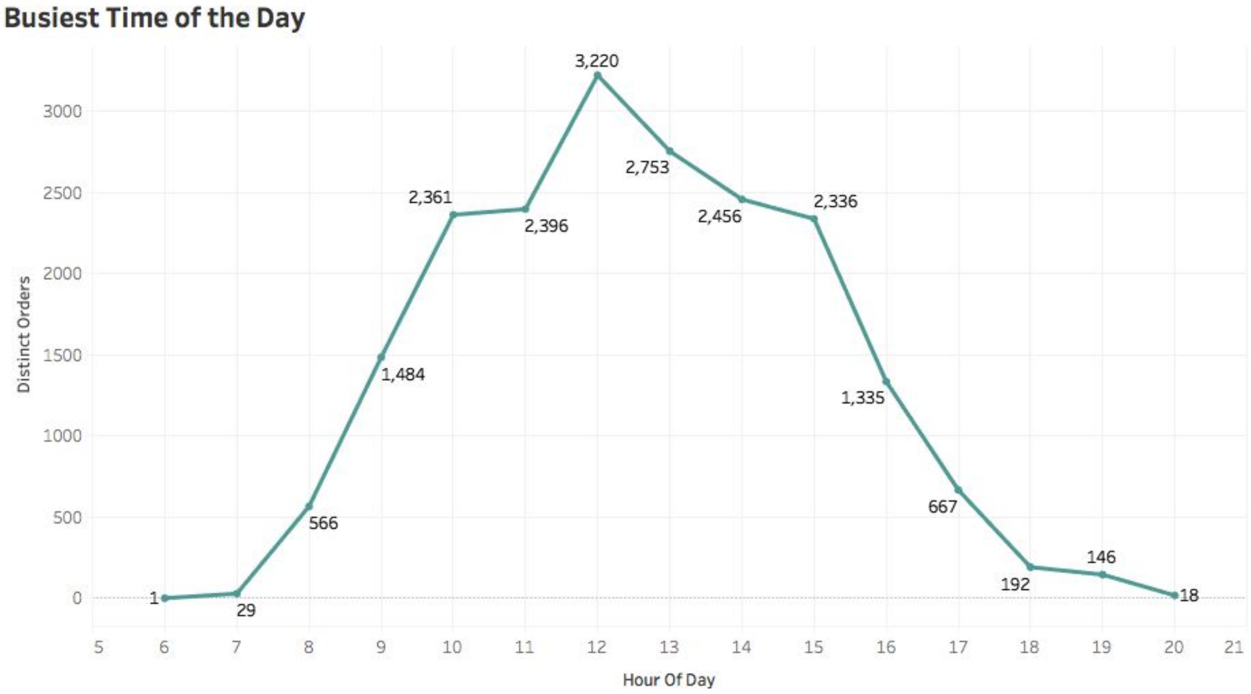
E. Busiest Days of the Week

There are no records for orders on a Saturday. Wednesday and Thursdays are the most popular days for orders.



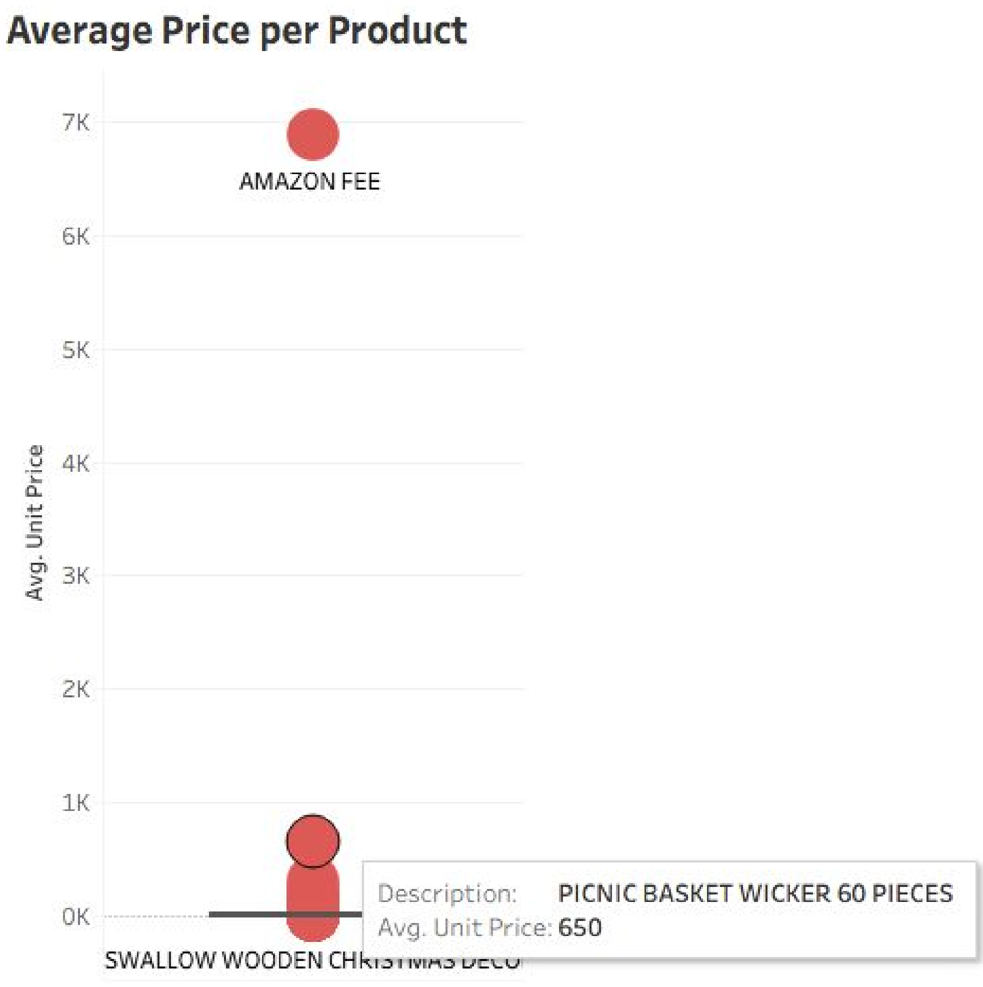
F. Busiest Hours of the Day

12 noon is when there is a serious spike in Orders. From 10am to 3pm are the busiest hours of the day.

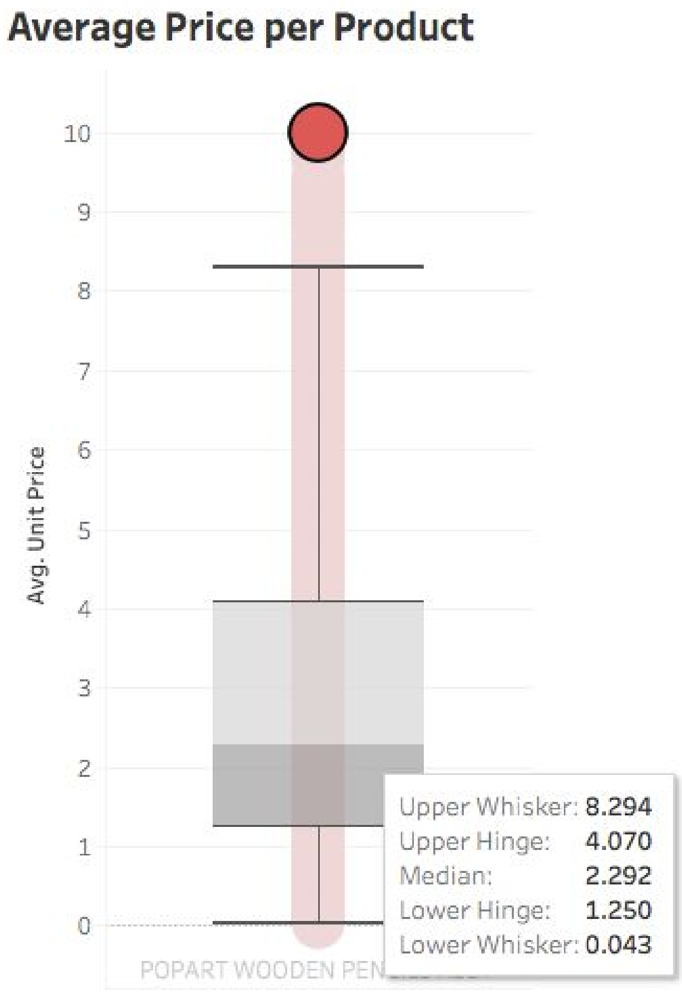


For an online business it may be important to ensure that the website is able to handle the number of customers are these peak hours of the day.

G. Some Average Price per Product Outliers



Filtering this diagram for Average Unit price < $10 provides a better view:



This shows that the majority of Products that the business sells are of low unit price, but there are a number of outliers that skew the data.

2. Summary/Notable Findings

- There are no recorded sales on any Saturdays in entire dataset - Is there a reason for this?
- There is a clear spike in Orders at 12pm - are stocks levels high at this time? Can our server handle a large number of visitors?
- November 2011 had the most Orders
- The most ordered Product has only 1 CustomerID
- ¼ of CustomerIDs are missing from the dataset - Can customers be segmented by this fact? Are these one-time customers?
- 2515 records had a UnitPrice = 0. Is there a business reason for this?
- The average price per Product is less than $3. There are some huge outliers with prices >$1000. These products could be further analysed.

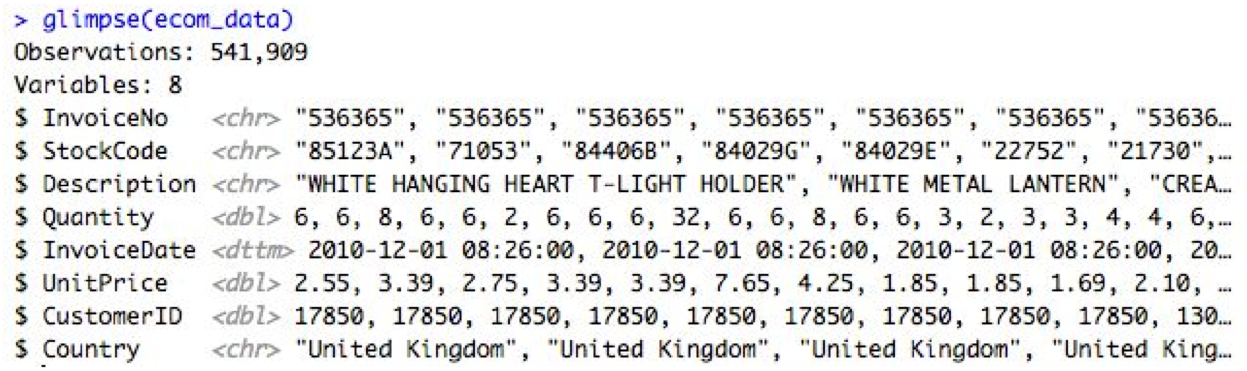
3. Data Pre-Processing

Firstly, opening the file in Excel allowed to see how it was laid out and gain a greater understanding.

It looked like an **Orders** database table.
There were **541909 records** for **8 fields**:

- InvoiceNo: *Order number*
- StockCode: *Item stock code*
- Description: *Item description*
- Quantity: *Quantity of item*
- InvoiceDate: *Date of order - with Day/Month written in USA format*
- UnitPrice: *Price of item*
- CustomerID: *Customer Identifier*
- Country: *Order country*

Then, loading the .csv file to R Studio.



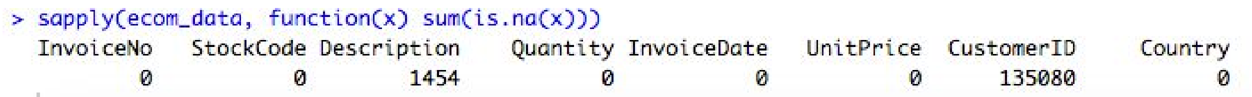
```
> glimpse(ecom_data)
Observations: 541,909
Variables: 8
$ InvoiceNo   <chr> "536365", "536365", "536365", "536365", "536365", "536365", "53636…
$ StockCode   <chr> "85123A", "71053", "84406B", "84029G", "84029E", "22752", "21730",…
$ Description <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHITE METAL LANTERN", "CREA…
$ Quantity    <dbl> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2, 3, 3, 4, 4, 6,…
$ InvoiceDate <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, 2010-12-01 08:26:00, 20…
$ UnitPrice   <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1.85, 1.69, 2.10, …
$ CustomerID  <dbl> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 130…
$ Country     <chr> "United Kingdom", "United Kingdom", "United Kingdom", "United King…
```

In interpreting the data presented - some business logic validation is important. E.g. Order Quantity and UnitPrice should be positive values, as well as checking for missing data.

Validation Checks

***Group 1**: Missing data/NA fields*

Checking for missing data per column:

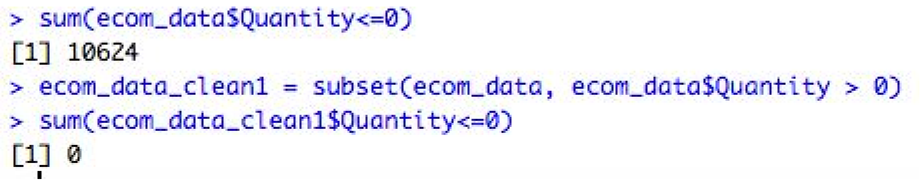


```
> sapply(ecom_data, function(x) sum(is.na(x)))
  InvoiceNo   StockCode Description    Quantity InvoiceDate   UnitPrice  CustomerID     Country 
          0           0        1454           0           0           0      135080           0 
```

There were 1454 records missing Description and 135,080 records missing CustomerID - removing all is ~25% of the dataset, which seemed like a lot so I decided to leave them for the moment and do some further analysis.

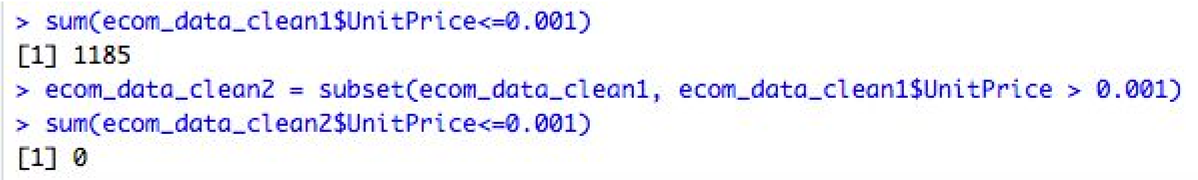
***Group 2:** Quantity less than or equal to 0*

I omitted any records in this group, as it does not make much business sense for this analysis.



```
> sum(ecom_data$Quantity<=0)
[1] 10624
> ecom_data_clean1 = subset(ecom_data, ecom_data$Quantity > 0)
> sum(ecom_data_clean1$Quantity<=0)
[1] 0
```

Group 3: *UnitPrice less than or equal to 0*

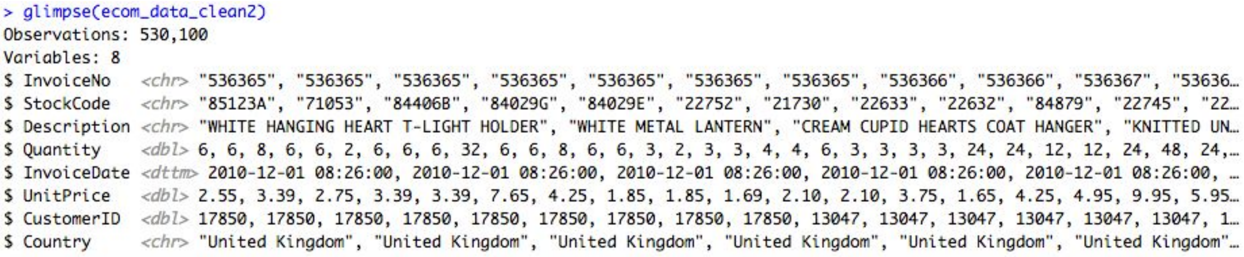


```
> sum(ecom_data_clean1$UnitPrice<=0.001)
[1] 1185
> ecom_data_clean2 = subset(ecom_data_clean1, ecom_data_clean1$UnitPrice > 0.001)
> sum(ecom_data_clean2$UnitPrice<=0.001)
[1] 0
```

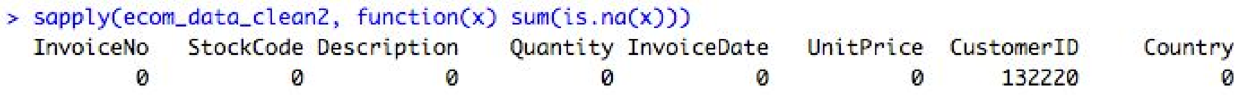
It is interesting that there are 2515 records where the unit price is equal to 0. I omitted them for this analysis but perhaps they could be used for a future analysis of how customers react to receiving ‘free’ items, if that is what they are.

Note - there was a crossover between records in Group 2 and 3.

The new dataset then contained 530,100 records - and no longer had any NA values for Description.



```
> glimpse(ecom_data_clean2)
Observations: 530,100
Variables: 8
$ InvoiceNo   <chr> "536365", "536365", "536365", "536365", "536365", "536365", "536365", "536366", "536366", "536367", "53636…
$ StockCode   <chr> "85123A", "71053", "84406B", "84029G", "84029E", "22752", "21730", "22633", "22632", "84879", "22745", "22…
$ Description <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHITE METAL LANTERN", "CREAM CUPID HEARTS COAT HANGER", "KNITTED UN…
$ Quantity    <dbl> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2, 3, 3, 4, 4, 6, 3, 3, 3, 3, 24, 24, 12, 12, 24, 48, 24,…
$ InvoiceDate <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, 2010-12-01 08:26:00, 2010-12-01 08:26:00, 2010-12-01 08:26:00, …
$ UnitPrice   <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1.85, 1.69, 2.10, 2.10, 3.75, 1.65, 4.25, 4.95, 9.95, 5.95…
$ CustomerID  <dbl> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 13047, 13047, 13047, 13047, 13047, 13047, 1…
$ Country     <chr> "United Kingdom", "United Kingdom", "United Kingdom", "United Kingdom", "United Kingdom", "United Kingdom"…
```

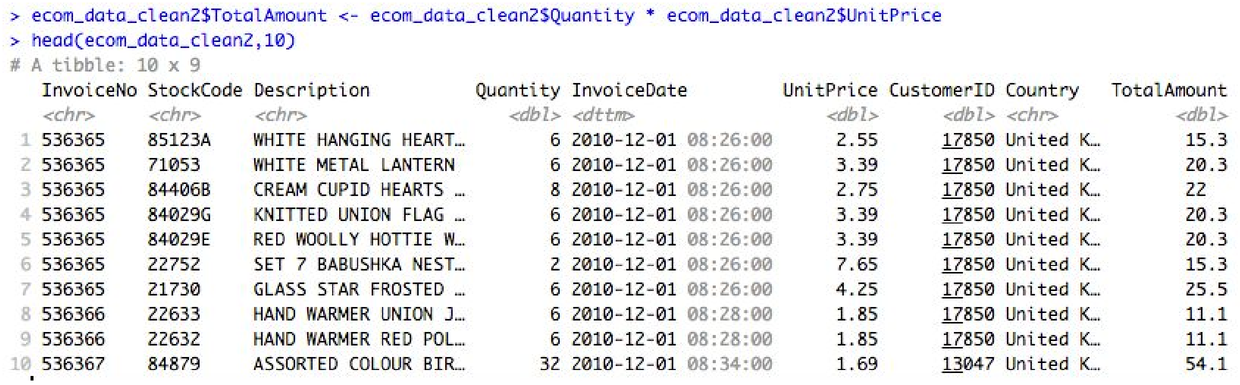


```
> sapply(ecom_data_clean2, function(x) sum(is.na(x)))
  InvoiceNo   StockCode Description    Quantity InvoiceDate   UnitPrice  CustomerID     Country 
          0           0           0           0           0           0      132220           0 
```

<u>Create New Fields</u>

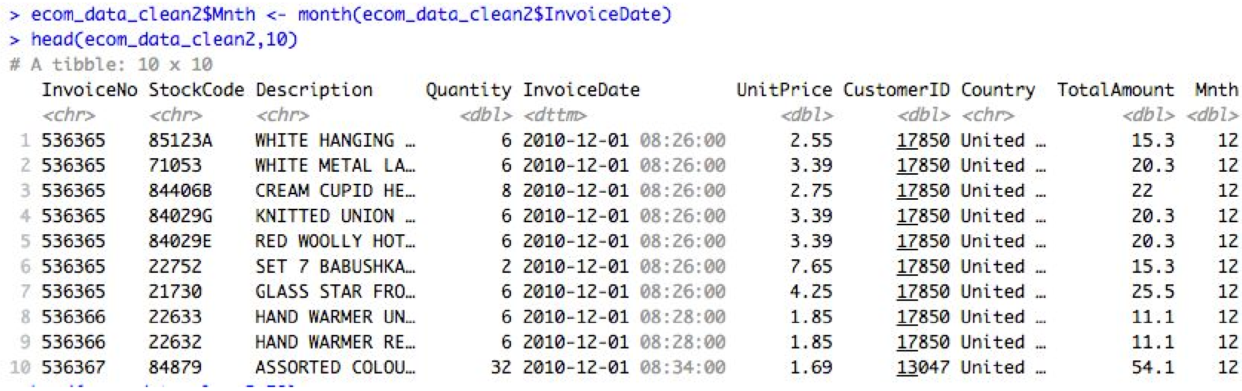
To make analysis easier I created 5 calculated fields.

TotalAmount: *Quantity * UnitPrice*

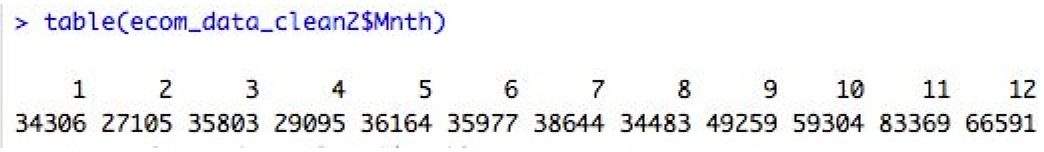


```
> ecom_data_clean2$TotalAmount <- ecom_data_clean2$Quantity * ecom_data_clean2$UnitPrice
> head(ecom_data_clean2,10)
# A tibble: 10 x 9
   InvoiceNo StockCode Description          Quantity InvoiceDate         UnitPrice CustomerID Country   TotalAmount
   <chr>     <chr>     <chr>                   <dbl> <dttm>                  <dbl>      <dbl> <chr>           <dbl>
 1 536365    85123A    WHITE HANGING HEART…        6 2010-12-01 08:26:00      2.55      17850 United K…        15.3
 2 536365    71053     WHITE METAL LANTERN         6 2010-12-01 08:26:00      3.39      17850 United K…        20.3
 3 536365    84406B    CREAM CUPID HEARTS …        8 2010-12-01 08:26:00      2.75      17850 United K…        22  
 4 536365    84029G    KNITTED UNION FLAG …        6 2010-12-01 08:26:00      3.39      17850 United K…        20.3
 5 536365    84029E    RED WOOLLY HOTTIE W…        6 2010-12-01 08:26:00      3.39      17850 United K…        20.3
 6 536365    22752     SET 7 BABUSHKA NEST…        2 2010-12-01 08:26:00      7.65      17850 United K…        15.3
 7 536365    21730     GLASS STAR FROSTED …        6 2010-12-01 08:26:00      4.25      17850 United K…        25.5
 8 536366    22633     HAND WARMER UNION J…        6 2010-12-01 08:28:00      1.85      17850 United K…        11.1
 9 536366    22632     HAND WARMER RED POL…        6 2010-12-01 08:28:00      1.85      17850 United K…        11.1
10 536367    84879     ASSORTED COLOUR BIR…       32 2010-12-01 08:34:00      1.69      13047 United K…        54.1
```

Month: *Parsed from InvoiceDate*

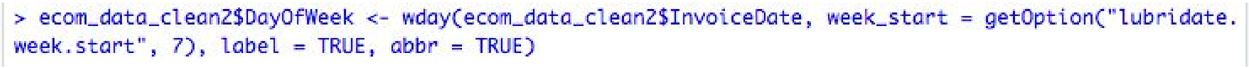


```
> ecom_data_clean2$Mnth <- month(ecom_data_clean2$InvoiceDate)
> head(ecom_data_clean2,10)
# A tibble: 10 x 10
   InvoiceNo StockCode Description     Quantity InvoiceDate         UnitPrice CustomerID Country  TotalAmount  Mnth
   <chr>     <chr>     <chr>              <dbl> <dttm>                  <dbl>      <dbl> <chr>          <dbl> <dbl>
 1 536365    85123A    WHITE HANGING …        6 2010-12-01 08:26:00      2.55      17850 United …        15.3    12
 2 536365    71053     WHITE METAL LA…        6 2010-12-01 08:26:00      3.39      17850 United …        20.3    12
 3 536365    84406B    CREAM CUPID HE…        8 2010-12-01 08:26:00      2.75      17850 United …        22      12
 4 536365    84029G    KNITTED UNION …        6 2010-12-01 08:26:00      3.39      17850 United …        20.3    12
 5 536365    84029E    RED WOOLLY HOT…        6 2010-12-01 08:26:00      3.39      17850 United …        20.3    12
 6 536365    22752     SET 7 BABUSHKA…        2 2010-12-01 08:26:00      7.65      17850 United …        15.3    12
 7 536365    21730     GLASS STAR FRO…        6 2010-12-01 08:26:00      4.25      17850 United …        25.5    12
 8 536366    22633     HAND WARMER UN…        6 2010-12-01 08:28:00      1.85      17850 United …        11.1    12
 9 536366    22632     HAND WARMER RE…        6 2010-12-01 08:28:00      1.85      17850 United …        11.1    12
10 536367    84879     ASSORTED COLOU…       32 2010-12-01 08:34:00      1.69      13047 United …        54.1    12
```

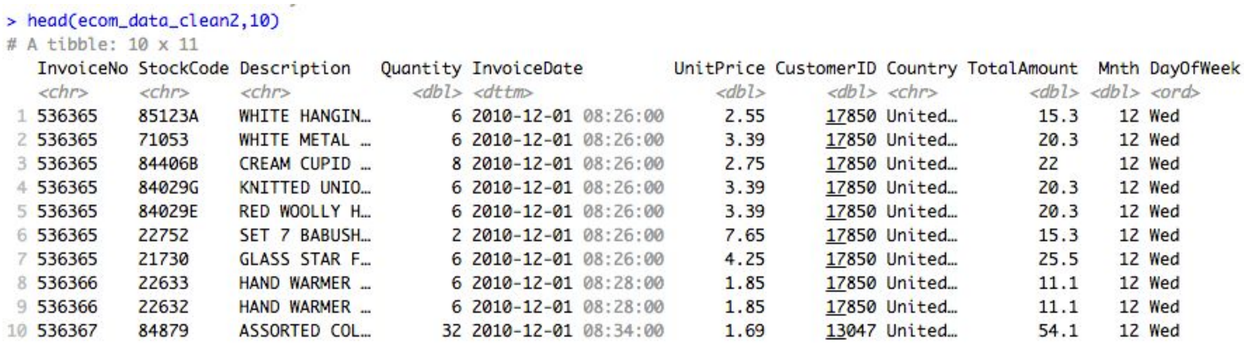


```
> table(ecom_data_clean2$Mnth)

    1     2     3     4     5     6     7     8     9    10    11    12 
34306 27105 35803 29095 36164 35977 38644 34483 49259 59304 83369 66591 
```

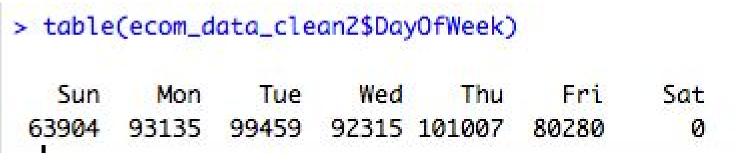
Day of the Week: *Parsed from InvoiceDate*



```
> ecom_data_clean2$DayOfWeek <- wday(ecom_data_clean2$InvoiceDate, week_start = getOption("lubridate.
week.start", 7), label = TRUE, abbr = TRUE)
```



```
> head(ecom_data_clean2,10)
# A tibble: 10 x 11
   InvoiceNo StockCode Description    Quantity InvoiceDate         UnitPrice CustomerID Country TotalAmount  Mnth DayOfWeek
   <chr>     <chr>     <chr>             <dbl> <dttm>                  <dbl>      <dbl> <chr>         <dbl> <dbl> <ord>    
 1 536365    85123A    WHITE HANGIN…         6 2010-12-01 08:26:00      2.55      17850 United…        15.3    12 Wed      
 2 536365    71053     WHITE METAL …         6 2010-12-01 08:26:00      3.39      17850 United…        20.3    12 Wed      
 3 536365    84406B    CREAM CUPID …         8 2010-12-01 08:26:00      2.75      17850 United…        22      12 Wed      
 4 536365    84029G    KNITTED UNIO…         6 2010-12-01 08:26:00      3.39      17850 United…        20.3    12 Wed      
 5 536365    84029E    RED WOOLLY H…         6 2010-12-01 08:26:00      3.39      17850 United…        20.3    12 Wed      
 6 536365    22752     SET 7 BABUSH…         2 2010-12-01 08:26:00      7.65      17850 United…        15.3    12 Wed      
 7 536365    21730     GLASS STAR F…         6 2010-12-01 08:26:00      4.25      17850 United…        25.5    12 Wed      
 8 536366    22633     HAND WARMER …         6 2010-12-01 08:28:00      1.85      17850 United…        11.1    12 Wed      
 9 536366    22632     HAND WARMER …         6 2010-12-01 08:28:00      1.85      17850 United…        11.1    12 Wed      
10 536367    84879     ASSORTED COL…        32 2010-12-01 08:34:00      1.69      13047 United…        54.1    12 Wed      
```

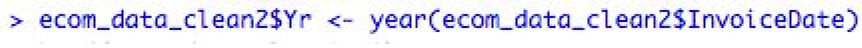


```
> table(ecom_data_clean2$DayOfWeek)

   Sun    Mon    Tue    Wed    Thu    Fri    Sat 
 63904  93135  99459  92315 101007  80280      0 
```

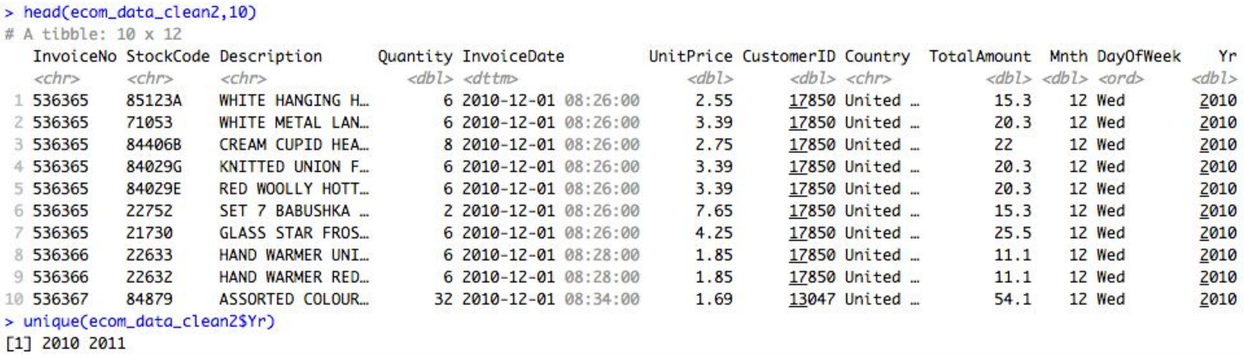
There was no orders on Saturdays which seems unusual.

It is assumed that the ordering system does not allow users to make orders on Saturday or that there is some missing data from this dataset.

Year: *Parsed from InvoiceDate*

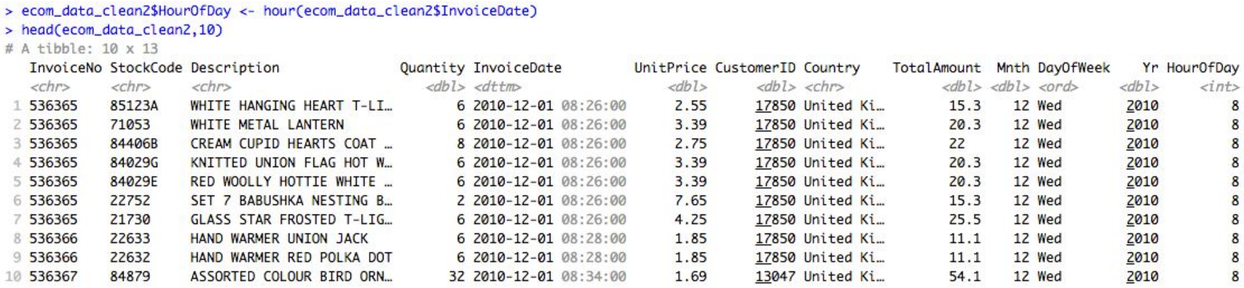


```
> ecom_data_clean2$Yr <- year(ecom_data_clean2$InvoiceDate)
```



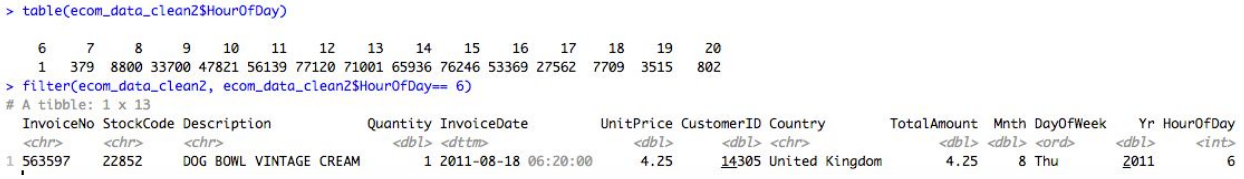
```
> head(ecom_data_clean2,10)
# A tibble: 10 x 12
   InvoiceNo StockCode Description       Quantity InvoiceDate         UnitPrice CustomerID Country  TotalAmount  Mnth DayOfWeek    Yr
   <chr>     <chr>     <chr>                <dbl> <dttm>                  <dbl>      <dbl> <chr>          <dbl> <dbl> <ord>     <dbl>
 1 536365    85123A    WHITE HANGING H…         6 2010-12-01 08:26:00      2.55      17850 United …        15.3    12 Wed        2010
 2 536365    71053     WHITE METAL LAN…         6 2010-12-01 08:26:00      3.39      17850 United …        20.3    12 Wed        2010
 3 536365    84406B    CREAM CUPID HEA…         8 2010-12-01 08:26:00      2.75      17850 United …        22      12 Wed        2010
 4 536365    84029G    KNITTED UNION F…         6 2010-12-01 08:26:00      3.39      17850 United …        20.3    12 Wed        2010
 5 536365    84029E    RED WOOLLY HOTT…         6 2010-12-01 08:26:00      3.39      17850 United …        20.3    12 Wed        2010
 6 536365    22752     SET 7 BABUSHKA …         2 2010-12-01 08:26:00      7.65      17850 United …        15.3    12 Wed        2010
 7 536365    21730     GLASS STAR FROS…         6 2010-12-01 08:26:00      4.25      17850 United …        25.5    12 Wed        2010
 8 536366    22633     HAND WARMER UNI…         6 2010-12-01 08:28:00      1.85      17850 United …        11.1    12 Wed        2010
 9 536366    22632     HAND WARMER RED…         6 2010-12-01 08:28:00      1.85      17850 United …        11.1    12 Wed        2010
10 536367    84879     ASSORTED COLOUR…        32 2010-12-01 08:34:00      1.69      13047 United …        54.1    12 Wed        2010
> unique(ecom_data_clean2$Yr)
[1] 2010 2011
```

Time of the Day: *Parsed from InvoiceDate*



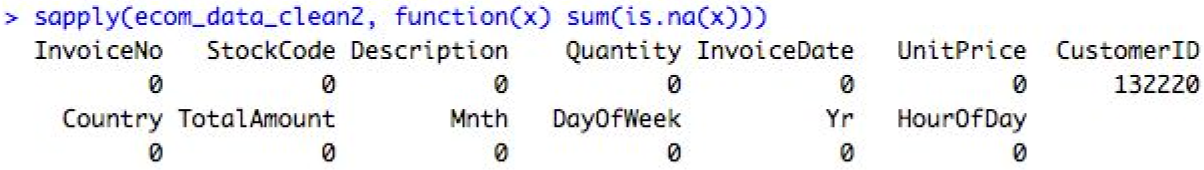
```
> ecom_data_clean2$HourOfDay <- hour(ecom_data_clean2$InvoiceDate)
> head(ecom_data_clean2,10)
# A tibble: 10 x 13
   InvoiceNo StockCode Description                Quantity InvoiceDate         UnitPrice CustomerID Country    TotalAmount  Mnth DayOfWeek    Yr HourOfDay
   <chr>     <chr>     <chr>                         <dbl> <dttm>                  <dbl>      <dbl> <chr>            <dbl> <dbl> <ord>     <dbl>     <int>
 1 536365    85123A    WHITE HANGING HEART T-LI…         6 2010-12-01 08:26:00      2.55      17850 United Ki…        15.3    12 Wed        2010         8
 2 536365    71053     WHITE METAL LANTERN               6 2010-12-01 08:26:00      3.39      17850 United Ki…        20.3    12 Wed        2010         8
 3 536365    84406B    CREAM CUPID HEARTS COAT …         8 2010-12-01 08:26:00      2.75      17850 United Ki…        22      12 Wed        2010         8
 4 536365    84029G    KNITTED UNION FLAG HOT W…         6 2010-12-01 08:26:00      3.39      17850 United Ki…        20.3    12 Wed        2010         8
 5 536365    84029E    RED WOOLLY HOTTIE WHITE …         6 2010-12-01 08:26:00      3.39      17850 United Ki…        20.3    12 Wed        2010         8
 6 536365    22752     SET 7 BABUSHKA NESTING B…         2 2010-12-01 08:26:00      7.65      17850 United Ki…        15.3    12 Wed        2010         8
 7 536365    21730     GLASS STAR FROSTED T-LIG…         6 2010-12-01 08:26:00      4.25      17850 United Ki…        25.5    12 Wed        2010         8
 8 536366    22633     HAND WARMER UNION JACK            6 2010-12-01 08:28:00      1.85      17850 United Ki…        11.1    12 Wed        2010         8
 9 536366    22632     HAND WARMER RED POLKA DOT         6 2010-12-01 08:28:00      1.85      17850 United Ki…        11.1    12 Wed        2010         8
10 536367    84879     ASSORTED COLOUR BIRD ORN…        32 2010-12-01 08:34:00      1.69      13047 United Ki…        54.1    12 Wed        2010         8
```

I checked that the HourOfDay field made sense by viewing all of the unique values. There was only 1 entry for 6AM - I filtered this record out to ensure the parsing had worked as intended. It looked good.



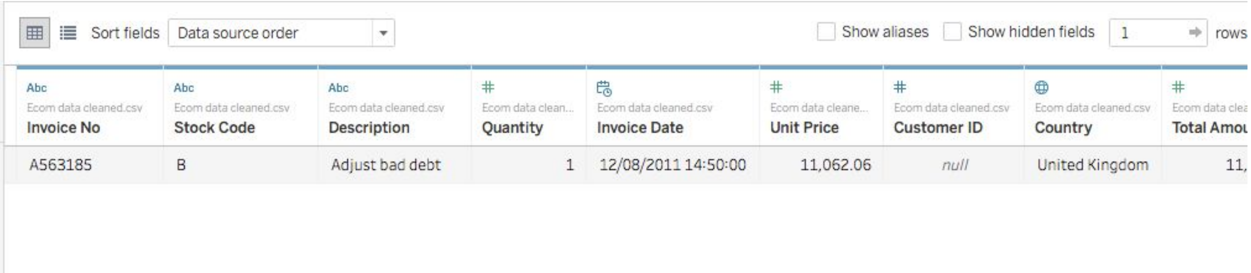
```
> table(ecom_data_clean2$HourOfDay)

    6     7     8     9    10    11    12    13    14    15    16    17    18    19    20 
    1   379  8800 33700 47821 56139 77120 71001 65936 76246 53369 27562  7709  3515   802 
> filter(ecom_data_clean2, ecom_data_clean2$HourOfDay== 6)
# A tibble: 1 x 13
  InvoiceNo StockCode Description           Quantity InvoiceDate         UnitPrice CustomerID Country        TotalAmount  Mnth DayOfWeek    Yr HourOfDay
  <chr>     <chr>     <chr>                    <dbl> <dttm>                  <dbl>      <dbl> <chr>                <dbl> <dbl> <ord>     <dbl>     <int>
1 563597    22852     DOG BOWL VINTAGE CREAM       1 2011-08-18 06:20:00      4.25      14305 United Kingdom        4.25     8 Thu        2011         6
```

The dataset then had 530,100 records and 13 variables. I rechecked any missing values. CustomerID remains the only field with missing values. This looked good.



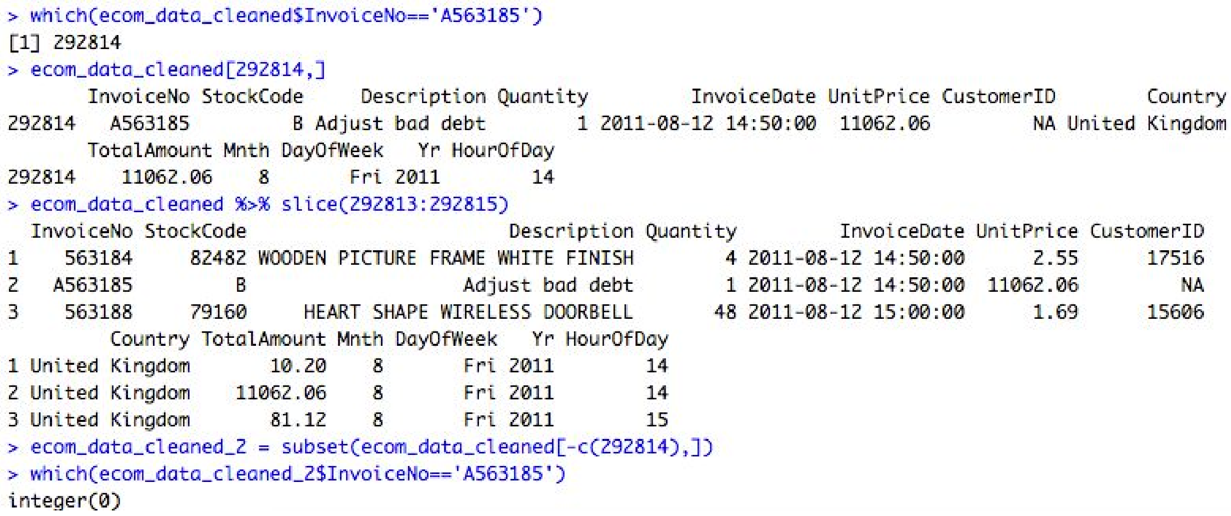
```
> sapply(ecom_data_clean2, function(x) sum(is.na(x)))
  InvoiceNo   StockCode Description    Quantity InvoiceDate   UnitPrice  CustomerID 
          0           0           0           0           0           0      132220 
    Country TotalAmount        Mnth   DayOfWeek          Yr   HourOfDay 
          0           0           0           0           0           0 
```

I tested loading this file set to Tableau. Tableau took ‘InvoiceNo’ as an Integer value. When I explored this further I realised that there was 1 record that contained a letter in the ‘InvoiceNo’



Abc Ecom data cleaned.csv Invoice No	Abc Ecom data cleaned.csv Stock Code	Abc Ecom data cleaned.csv Description	# Ecom data clean... Quantity	Ecom data cleaned.csv Invoice Date	# Ecom data cleane... Unit Price	# Ecom data cleaned.csv Customer ID	Ecom data cleaned.csv Country	# Ecom data clea Total Amou
A563185	B	Adjust bad debt	1	12/08/2011 14:50:00	11,062.06	null	United Kingdom	11,

I decided to remove this record - based on the ‘Adjusted bad debt’ Description field, it did not seem to fit in with the other records. I removed this in R Studio as seen below:



```
> which(ecom_data_cleaned$InvoiceNo=='A563185')
[1] 292814
> ecom_data_cleaned[292814,]
       InvoiceNo StockCode     Description Quantity         InvoiceDate UnitPrice CustomerID        Country
292814   A563185         B Adjust bad debt        1 2011-08-12 14:50:00  11062.06         NA United Kingdom
       TotalAmount Mnth DayOfWeek   Yr HourOfDay
292814    11062.06    8       Fri 2011        14
> ecom_data_cleaned %>% slice(292813:292815)
  InvoiceNo StockCode                        Description Quantity         InvoiceDate UnitPrice CustomerID
1    563184     82482 WOODEN PICTURE FRAME WHITE FINISH        4 2011-08-12 14:50:00      2.55      17516
2   A563185         B                    Adjust bad debt        1 2011-08-12 14:50:00  11062.06         NA
3    563188     79160      HEART SHAPE WIRELESS DOORBELL       48 2011-08-12 15:00:00      1.69      15606
         Country TotalAmount Mnth DayOfWeek   Yr HourOfDay
1 United Kingdom       10.20    8       Fri 2011        14
2 United Kingdom    11062.06    8       Fri 2011        14
3 United Kingdom       81.12    8       Fri 2011        15
> ecom_data_cleaned_2 = subset(ecom_data_cleaned[-c(292814),])
> which(ecom_data_cleaned_2$InvoiceNo=='A563185')
integer(0)
```

Finally, I loaded the clean dataset in Tableau - it contained 530099 records.