AIT 580 - Big Data Analytics Final Project

Ayushi Tiwari

George Mason University

Author Note

Ayushi Tiwari, Data Analytics Engineering, Volgenau School of Engineering

Contact: atiwari4@masonlive.gmu.edu

Abstract

This paper explores various visualization techniques to analyze a fairly large dataset to answer specific questions. It employs several visualization and data management techniques studied during the course. It further attempts to conclude by providing advance analysis techniques that can be employed to further perform informed decisions.

Keywords: Visualization Techniques, Data Analysis

Introduction

This project depicts the learning from AIT-580 course. The objective is to use a variety of data analysis techniques on the chosen fairly large dataset to analyze and interpret results for decision making.

The data selected for this project are the online transactions from an E-Commerce website in United Kingdom for the year 2011. The original dataset contains around 550,000 transaction records of all occasion gift items sold on the website.

A number of data visualization techniques were employed on the dataset to identify top selling products, top customers, weekly and monthly revenue and sales in different countries of the world. Techniques such as Time series and Linear Regression were employed for in –depth data analysis.

Who

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.

The source of the dataset is http://archive.ics.uci.edu/ml/datasets/online+retail.

Need

This is an interesting dataset of sales from an e-commerce website. It was necessary to collect this data in order to use it improve sales as well as customer experience. There is scope for good analysis and useful data decisions can be made. It is well known that e-commerce is a buzzing trend in the market these days and has gained immense popularity in the last decade. This is evident from the huge number of transactions in the dataset occurring almost every minute. A number of analyses can be performed on the dataset. Several potential questions can be answered through this data:

1. What is the sales trend on particular days of the week as well as month in a year?
2. Which are the top selling products and top buying customers?
3. What is the revenue trend over a period of one year?
4. Is there a relationship between unit price of a product and the quantity of product purchased?

The original dataset needed a lot of cleansing as there were some bad data rows that weren’t useful in the analysis.

Data Cleansing

Following measures were taken to clean up the original dataset:

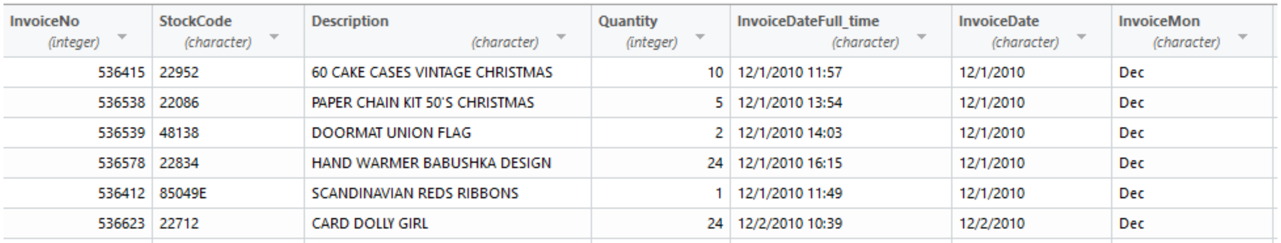
1. Records containing **negative values in the Quantity column** were removed as they represented the cancelled orders.
2. Records with **missing customer ID** were removed.
3. There were records identified with **unreasonably high Unit price**. These were Amazon fee, Bank Charges, Postage charges and Manual Fees. All such records were removed.
4. Records with **zero Unit Price** were also deleted.
5. There were records identified with similar Stock Id but slightly different Description. These records were updated to use a single unique description for each Stock Id.

Data Enrichment

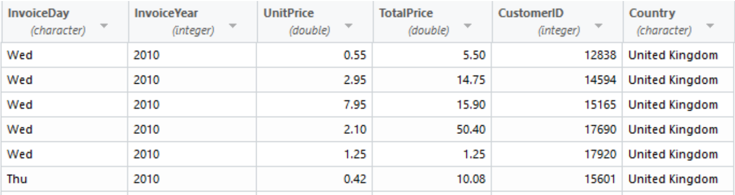
In order to perform a variety of analysis, some columns were added to the existing dataset. The date and the month part were extracted into separate fields from the InvoiceDate field. Day of the week field was deduced from the InvoiceDate field. A new field of "TotalPrice" was added which was calculated by the product of UnitPrice and Quantity Field.

Data Sampling

There were approximately 400,000 records left after performing data cleansing as discussed above. However, it is still a huge dataset and would require higher RAM capability to perform analysis using RStudio. Hence, the dataset was evenly sampled and narrow down to 100,000 rows to perform rows. Ranking technique was employed to extract 1 record out of every 4 records in the same chronological order. All the analysis was performed in this sample set of 100,000 records.



InvoiceNo (integer)	StockCode (character)	Description (character)	Quantity (integer)	InvoiceDateFull_time (character)	InvoiceDate (character)	InvoiceMon (character)
536415	22952	60 CAKE CASES VINTAGE CHRISTMAS	10	12/1/2010 11:57	12/1/2010	Dec
536538	22086	PAPER CHAIN KIT 50'S CHRISTMAS	5	12/1/2010 13:54	12/1/2010	Dec
536539	48138	DOORMAT UNION FLAG	2	12/1/2010 14:03	12/1/2010	Dec
536578	22834	HAND WARMER BABUSHKA DESIGN	24	12/1/2010 16:15	12/1/2010	Dec
536412	85049E	SCANDINAVIAN REDS RIBBONS	1	12/1/2010 11:49	12/1/2010	Dec
536623	22712	CARD DOLLY GIRL	24	12/2/2010 10:39	12/2/2010	Dec



InvoiceDay (character)	InvoiceYear (integer)	UnitPrice (double)	TotalPrice (double)	CustomerID (integer)	Country (character)
Wed	2010	0.55	5.50	12838	United Kingdom
Wed	2010	2.95	14.75	14594	United Kingdom
Wed	2010	7.95	15.90	15165	United Kingdom
Wed	2010	2.10	50.40	17690	United Kingdom
Wed	2010	1.25	1.25	17920	United Kingdom
Thu	2010	0.42	10.08	15601	United Kingdom

Figure 1 Sample Dataset

Requirements and Resources Needed

The dataset cleaning was performed using both Excel and RStudio. Most of the analysis was performed using RStudio. A lot of websites and articles were studied in order to finalize the data. System requirements were matched in order work on such a large dataset.

Dataset Schema

There is a variety of variables in the dataset such as both categorical as well as continuous variables are present. Following is the schema of the selected dataset:

1. **InvoiceNo**: Invoice number. A 6-digit integral number uniquely assigned to each transaction.
2. **StockCode**: Product code. A 5-digit integral number uniquely assigned to each distinct product.
3. **Description**: Product name.
4. **Quantity**: The quantities of each product per transaction.
5. **InvoiceDate**: Invoice Date and time. The day and time when each transaction was generated.
6. **UnitPrice**: Unit price. Product price per unit in sterling.
7. **CustomerID**: Customer number. A 5-digit integral number uniquely assigned to each customer.
8. **Country**: Country name. The name of the country where each customer resides.

Results/Findings

Revenue by Month (Year 2011)

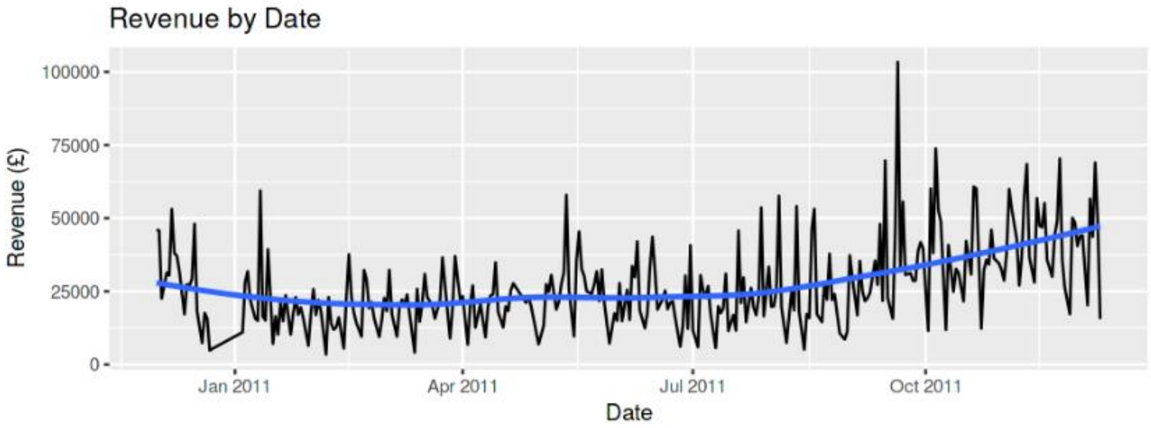


Figure 2 Time series plot for revenue

The Time series plot in Figure 2 depicts monthly revenue generated from sales in the year 2011. It can be clearly seen that sales rose considerably in the period Sep-Nov. This is the time when festive season is approaching and evidently people wish to send gifts to their friends and families. However, the sales seem to drop in the month of December. This could be because we do not have complete data for the month of December.

Transactions during a day

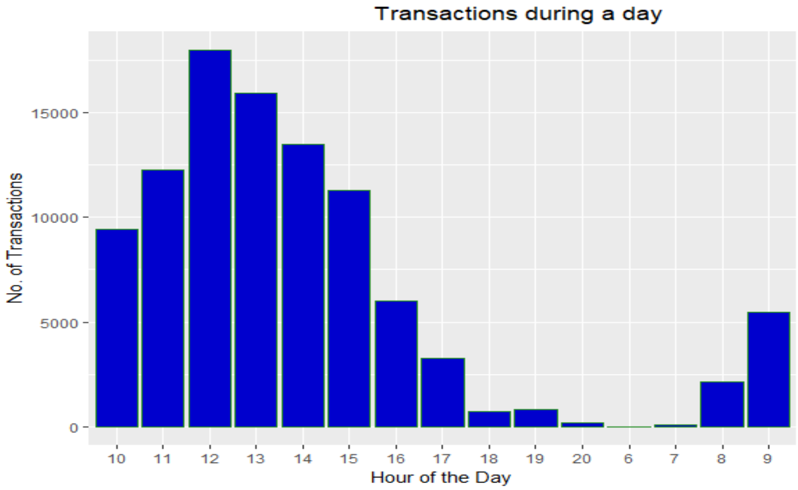


Figure 3 Transactions during a day

Figure 3 depicts number of transactions on hourly basis. Transactions seem to rise from 10 am to 12 pm and seem to drop thereafter.

Revenue by Day of the Week

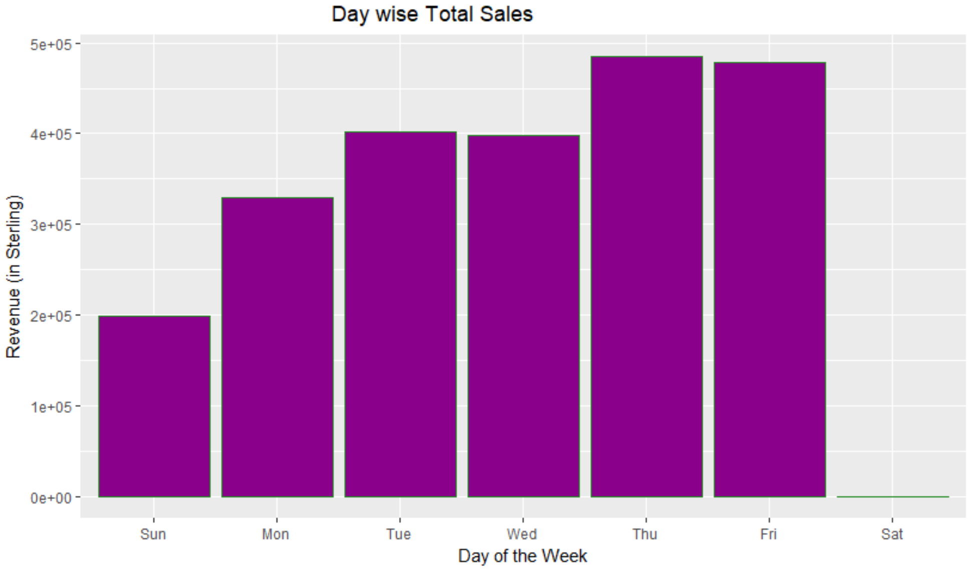


Figure 3 Revenue (in Sterling) by day of the week

The bar graph in Figure 3 represents the sales by days of the week. It is interesting to note that how sales seem to rise from Sundays till Friday with Thursdays showing highest purchases by customers during the week. No sales were reported on Saturday.

Top Selling Products

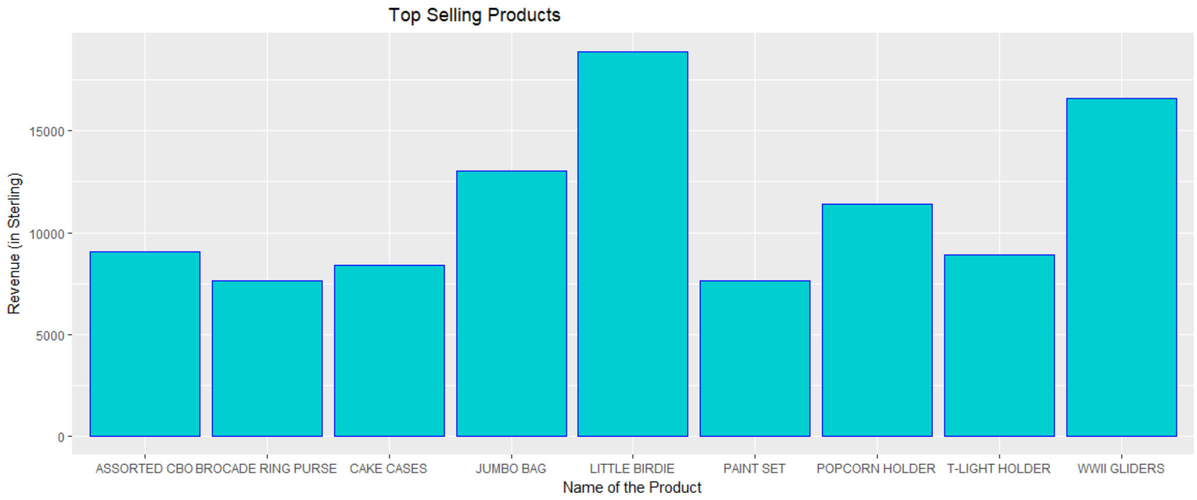


Figure 4 Top selling Products

Figure 4 shows the top selling products with “LITTLE BIRDIE” being the highest selling product followed by “WWII GLIDERS” and “JUMBO BAG”.

Top Buying Customers

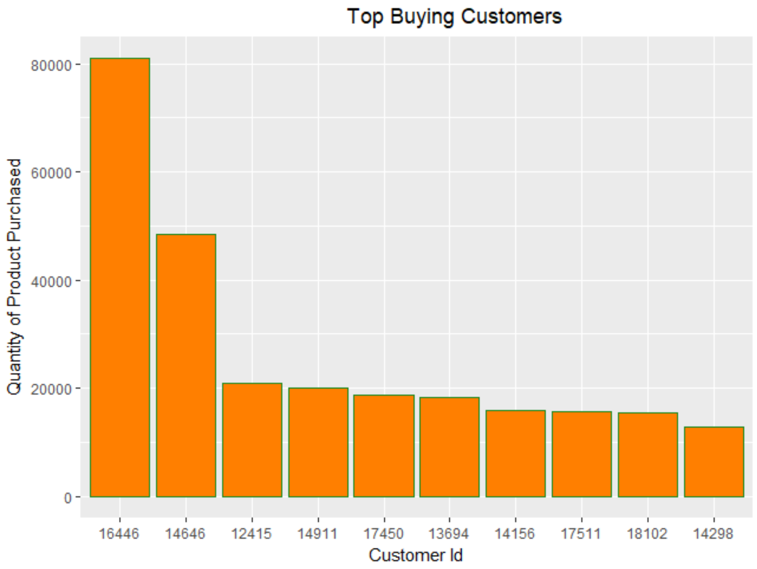


Figure 5 Top buying customers

Figure 5 depicts top 10 customers who have made the highest purchases in the span of one year (2011). Customer ID 15446 is the customer who purchased around 8000 quantity of products. Looks like he is a wholesale seller.

Linear Regression Model

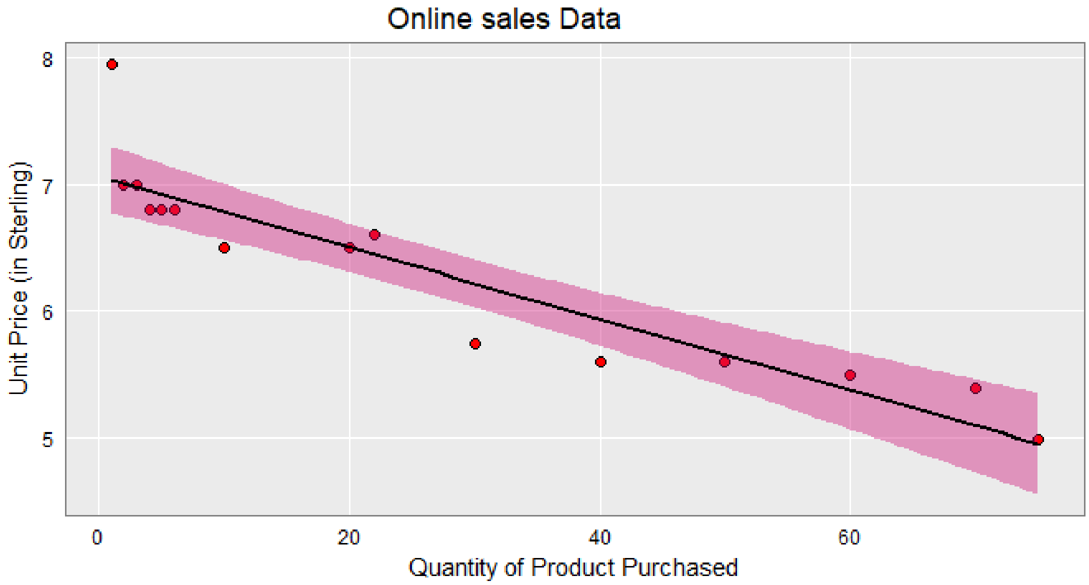
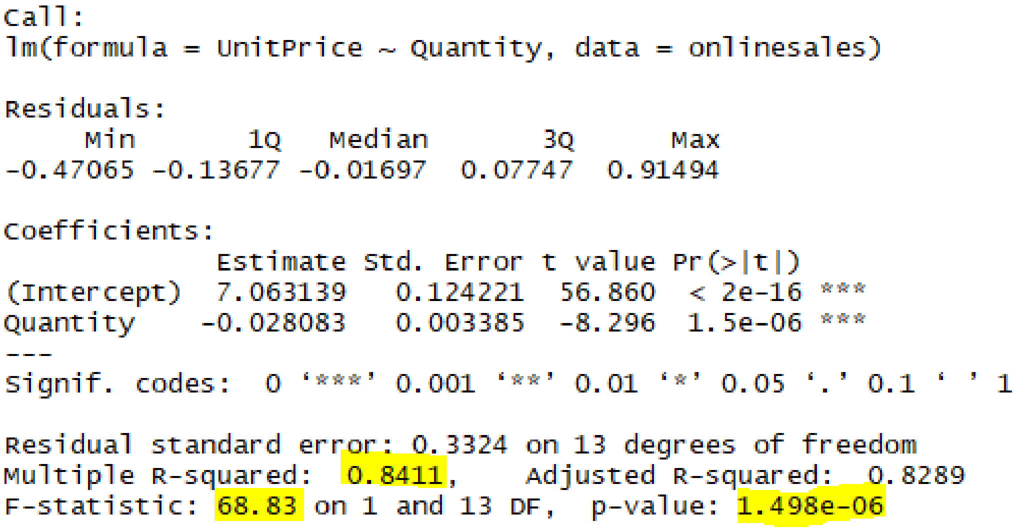


Figure 6 Declining trend in Unit price of Product

Regression Analysis Summary



```
Call:
lm(formula = UnitPrice ~ Quantity, data = onlinesales)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.47065 -0.13677 -0.01697  0.07747  0.91494 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.063139   0.124221  56.860  < 2e-16 ***
Quantity    -0.028083   0.003385  -8.296  1.5e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3324 on 13 degrees of freedom
Multiple R-squared:  0.8411,	Adjusted R-squared:  0.8289 
F-statistic: 68.83 on 1 and 13 DF,  p-value: 1.498e-06
```

The Regression plot has a negative slope showing a declining trend. The R-Square statistic of 0.84 represents the percent of the total variation in the dependent variable that is explained by the independent variables, i.e., the model's overall goodness of fit. So, 84% variability is explained by the model. The p-value of 1.498 x 10^{-9} is also highly significant. Hence, overall the model achieved through regression analysis looks significant with a good F – Statistic value.

Conclusion

An in-depth analysis was performed using R –Programming and various methods learnt in AIT-580 class. A variety of visualization techniques such as Bar Plots, Regression model, time series etc. were employed. This can help decision makers to increase the productivity and make informed and profitable decisions.

Predicting sales trend help us to know about the products whose stock we may consider to increase or reduce. Identifying top most customers help us to build better relationships with them and maybe provide them with discounts and good deals to consider. Shiny is an impressive tool to tell the data story in an interactive format lets other interact with the data and its analysis.

This dataset offers myriad opportunities for practicing skills in e-commerce sales analysis and customer segmentation. There are some other variables it would be nice to have in the dataset such as categories for the products. Additionally, before performing analysis it would be important to talk with the e-commerce team to understand the business and its customers and its strategic and tactical objectives. Knowing what the business wants to achieve, and what questions it has are central to performing a relevant analysis that generates actionable insight.

References

Chen, D. Online Retail Dataset (2015). Retrieved from http://archive.ics.uci.edu/ml/datasets/online+retail

Prabhakaran, S. Linear Regression. Retrieved from

http://r-statistics.co/Linear-Regression.html

Laney, D (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety.

Retrieved from

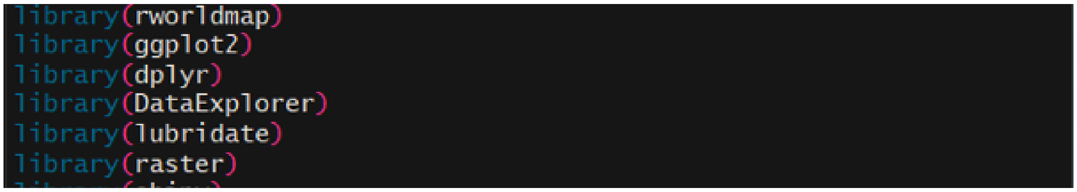
https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-DataManagement-Controlling-Data-Volume-Velocity-and-Variety.pdf

NIST BIG DATA Interoperability Framework: VOLUME 1, Definitions (2015) (p.4,5) .

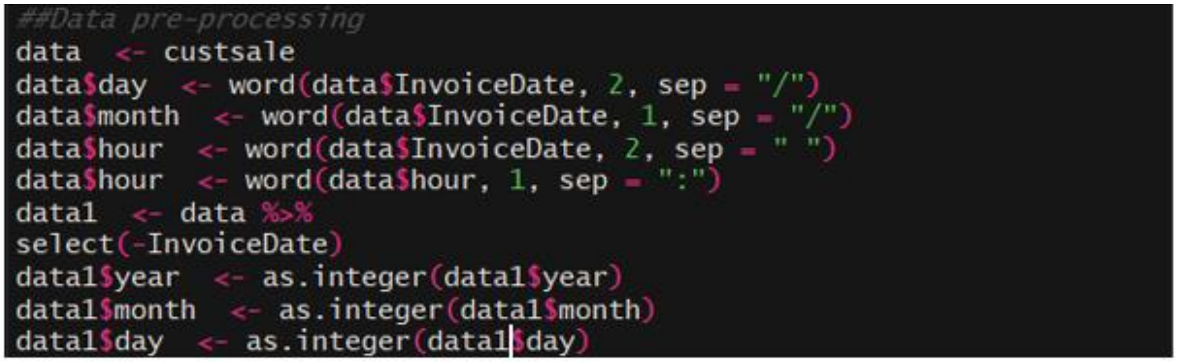
Retrieved from

https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

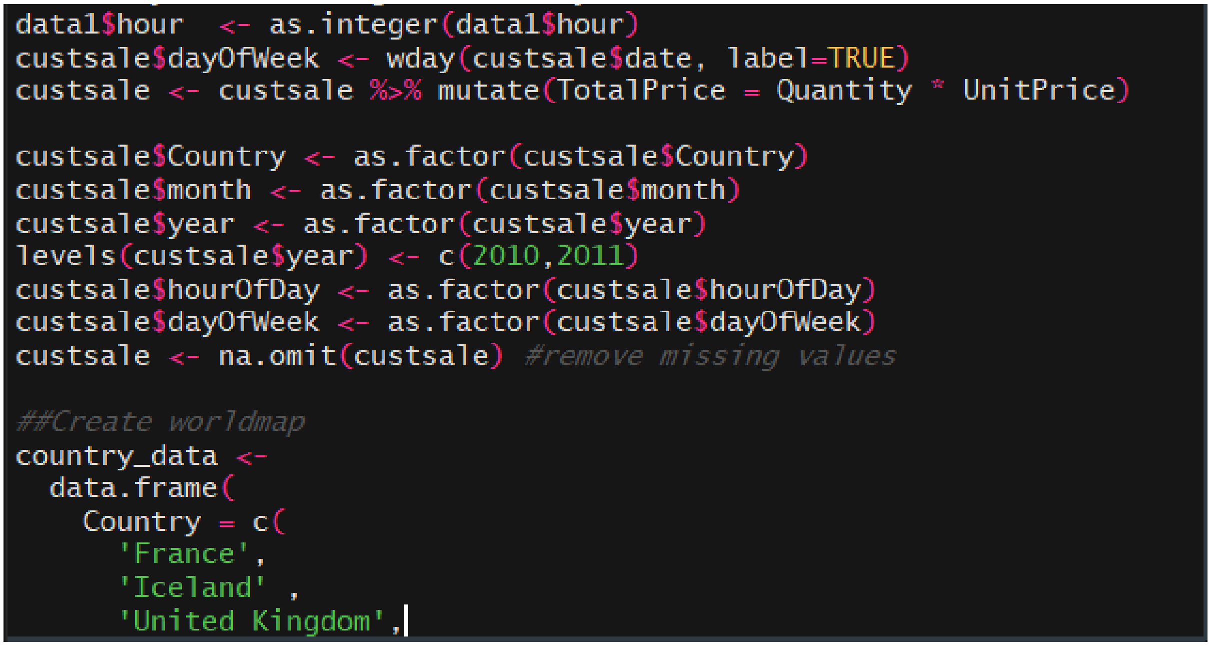
APPENDIX - RStudio code



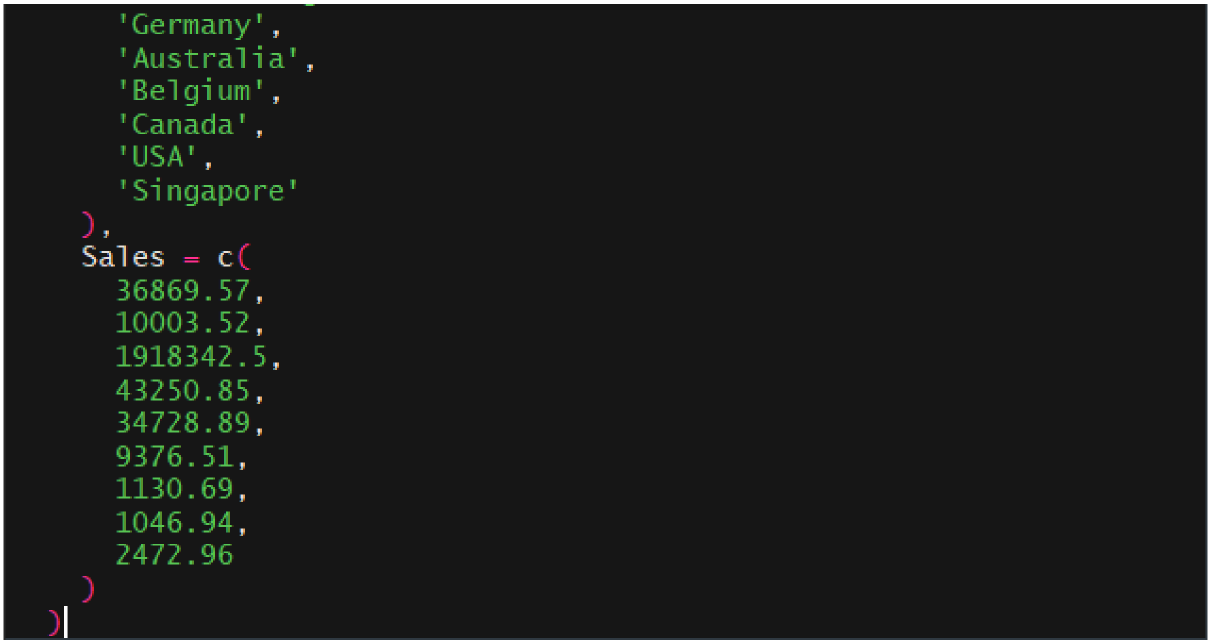
```r
library(rworldmap)
library(ggplot2)
library(dplyr)
library(DataExplorer)
library(lubridate)
library(raster)
```



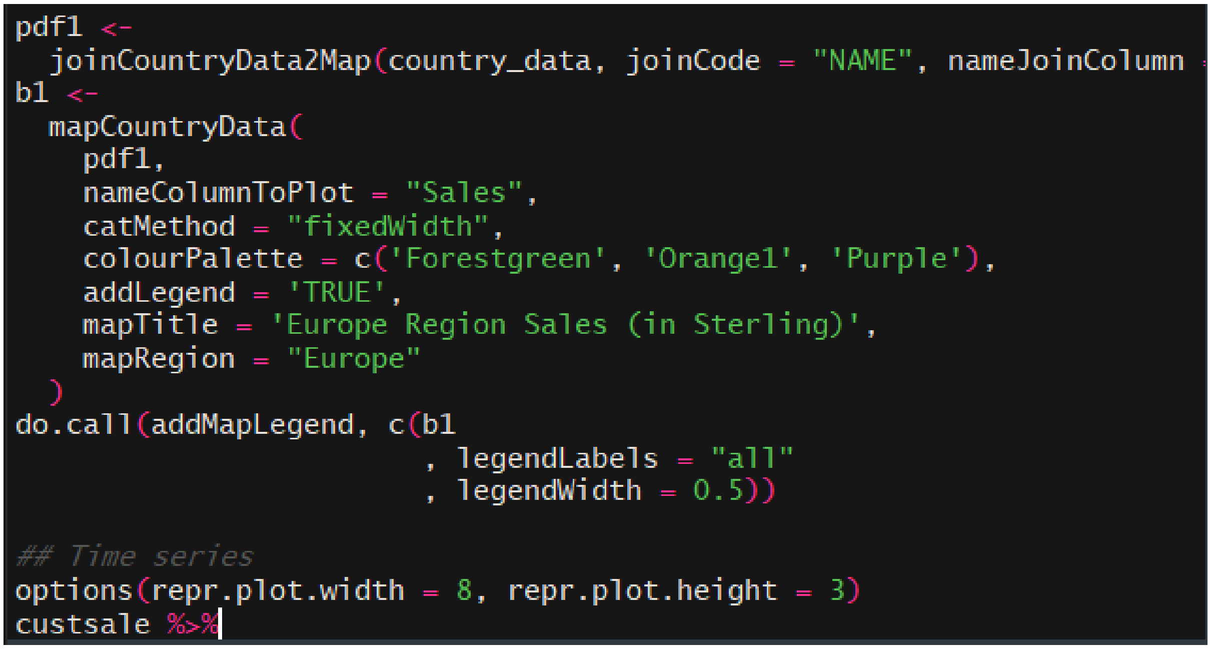
```r
##Data pre-processing
data  <- custsale
data$day  <- word(data$InvoiceDate, 2, sep = "/")
data$month  <- word(data$InvoiceDate, 1, sep = "/")
data$hour  <- word(data$InvoiceDate, 2, sep = " ")
data$hour  <- word(data$hour, 1, sep = ":")
data1  <- data %>%
select(-InvoiceDate)
data1$year  <- as.integer(data1$year)
data1$month  <- as.integer(data1$month)
data1$day  <- as.integer(data1$day)
```



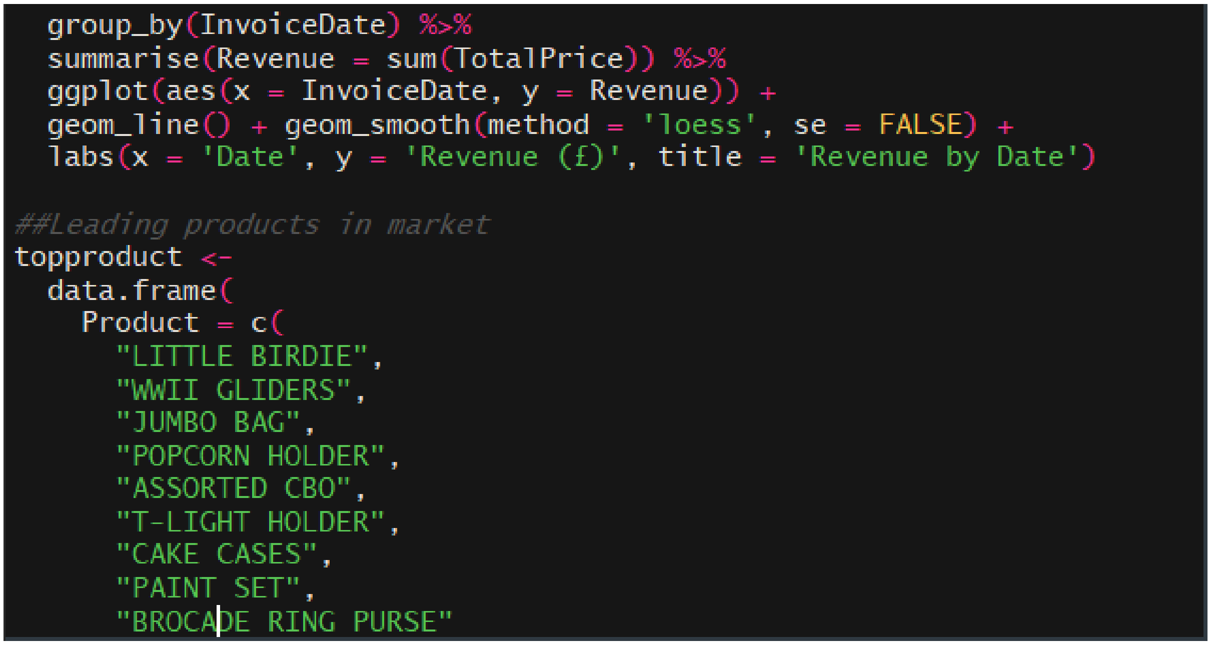
```r
data1$hour  <- as.integer(data1$hour)
custsale$dayOfWeek <- wday(custsale$date, label=TRUE)
custsale <- custsale %>% mutate(TotalPrice = Quantity * UnitPrice)

custsale$Country <- as.factor(custsale$Country)
custsale$month <- as.factor(custsale$month)
custsale$year <- as.factor(custsale$year)
levels(custsale$year) <- c(2010,2011)
custsale$hourOfDay <- as.factor(custsale$hourOfDay)
custsale$dayOfWeek <- as.factor(custsale$dayOfWeek)
custsale <- na.omit(custsale) #remove missing values

##Create worldmap
country_data <-
  data.frame(
    Country = c(
      'France',
      'Iceland' ,
      'United Kingdom',
```



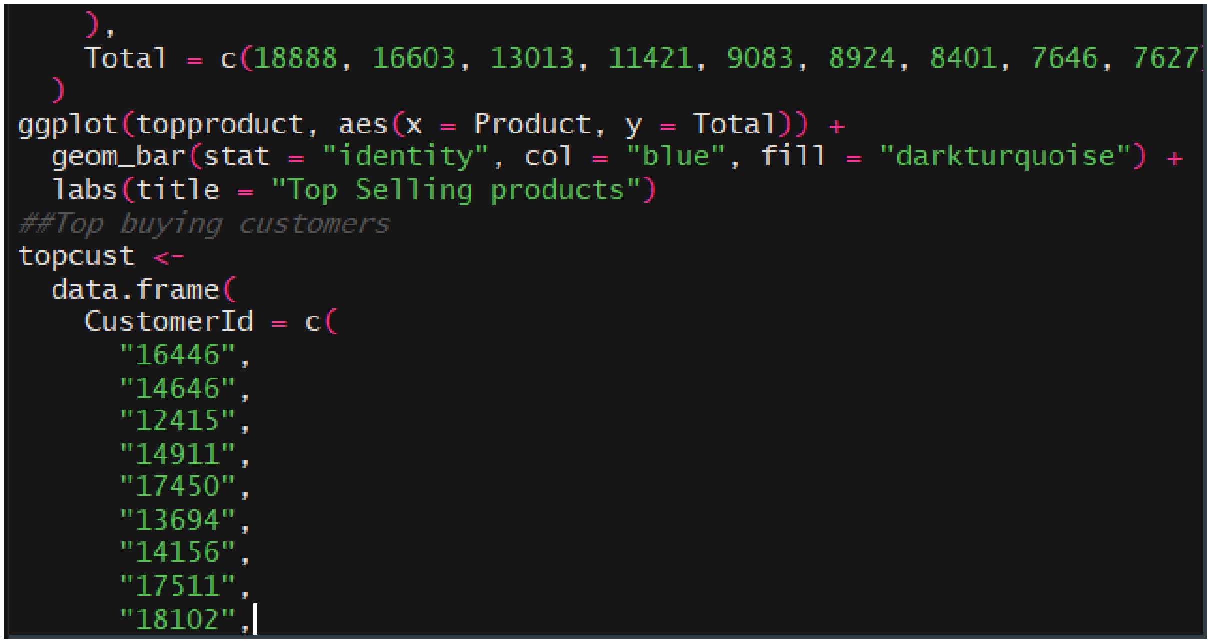
```
      'Germany',
      'Australia',
      'Belgium',
      'Canada',
      'USA',
      'Singapore'
    ),
    Sales = c(
      36869.57,
      10003.52,
      1918342.5,
      43250.85,
      34728.89,
      9376.51,
      1130.69,
      1046.94,
      2472.96
    )
  )
```



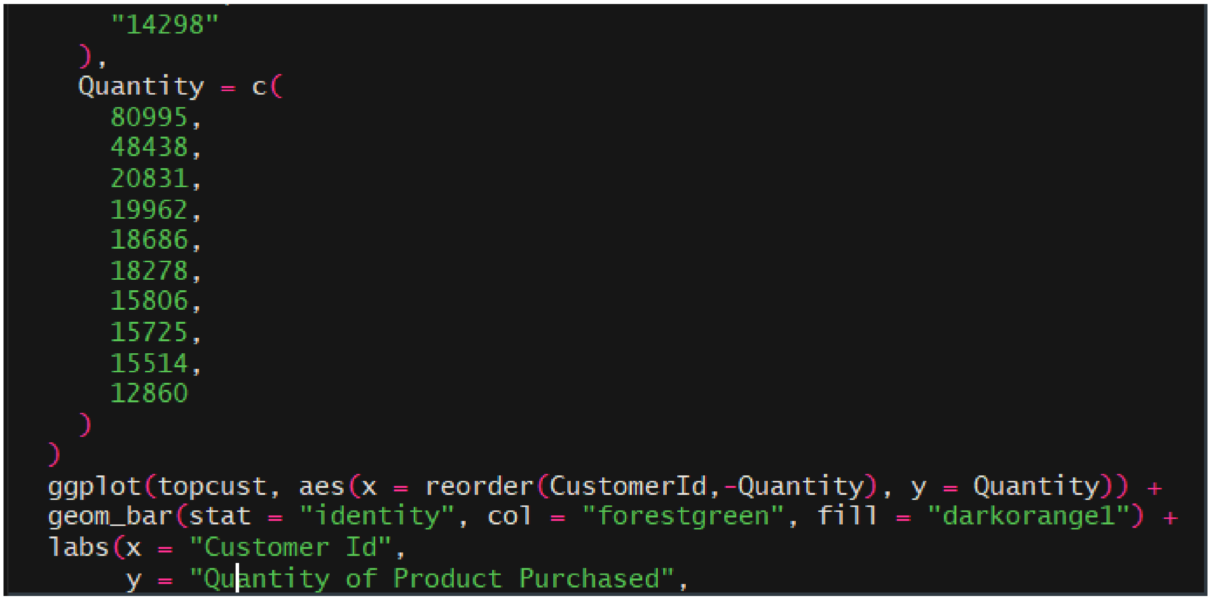
```
pdf1 <-
  joinCountryData2Map(country_data, joinCode = "NAME", nameJoinColumn
b1 <-
  mapCountryData(
    pdf1,
    nameColumnToPlot = "Sales",
    catMethod = "fixedWidth",
    colourPalette = c('Forestgreen', 'Orange1', 'Purple'),
    addLegend = 'TRUE',
    mapTitle = 'Europe Region Sales (in Sterling)',
    mapRegion = "Europe"
  )
do.call(addMapLegend, c(b1
                        , legendLabels = "all"
                        , legendWidth = 0.5))

## Time series
options(repr.plot.width = 8, repr.plot.height = 3)
custsale %>%
```



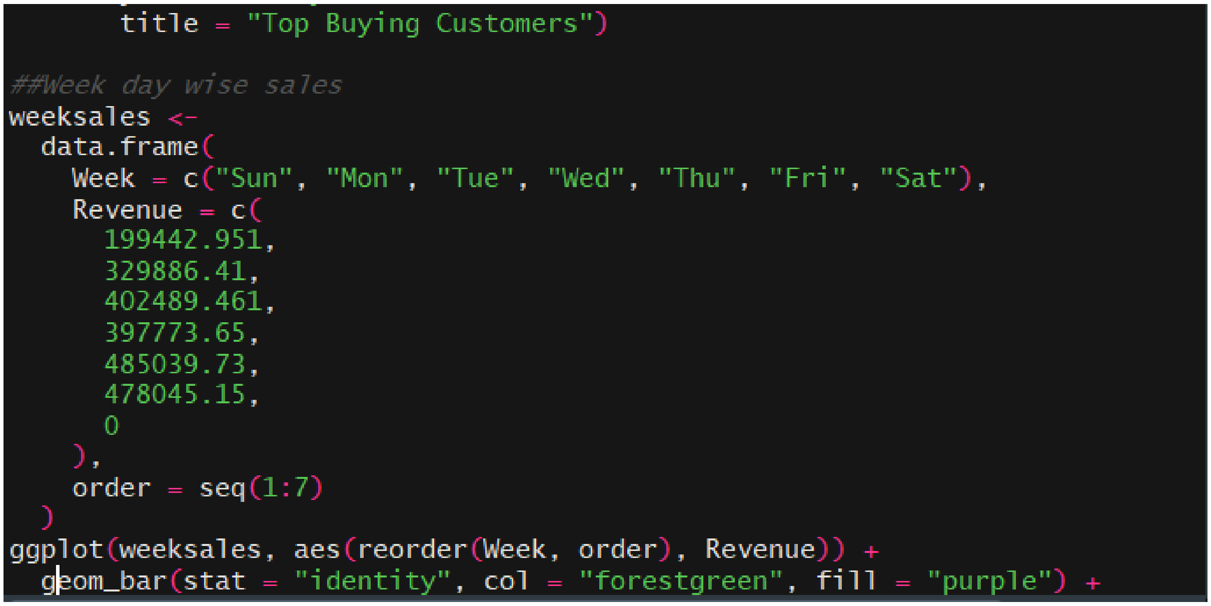
```
  group_by(InvoiceDate) %>%
  summarise(Revenue = sum(TotalPrice)) %>%
  ggplot(aes(x = InvoiceDate, y = Revenue)) +
  geom_line() + geom_smooth(method = 'loess', se = FALSE) +
  labs(x = 'Date', y = 'Revenue (£)', title = 'Revenue by Date')

##Leading products in market
topproduct <-
  data.frame(
    Product = c(
      "LITTLE BIRDIE",
      "WWII GLIDERS",
      "JUMBO BAG",
      "POPCORN HOLDER",
      "ASSORTED CBO",
      "T-LIGHT HOLDER",
      "CAKE CASES",
      "PAINT SET",
      "BROCADE RING PURSE"
```



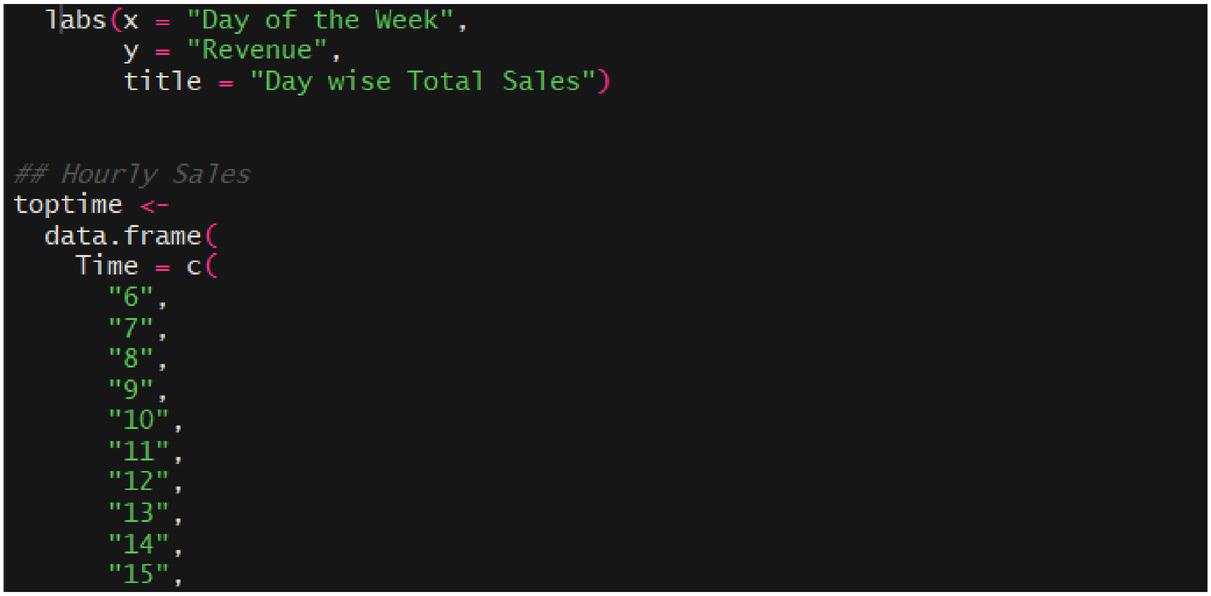
```
    ),
    Total = c(18888, 16603, 13013, 11421, 9083, 8924, 8401, 7646, 7627
  )
ggplot(topproduct, aes(x = Product, y = Total)) +
  geom_bar(stat = "identity", col = "blue", fill = "darkturquoise") +
  labs(title = "Top Selling products")
##Top buying customers
topcust <-
  data.frame(
    CustomerId = c(
      "16446",
      "14646",
      "12415",
      "14911",
      "17450",
      "13694",
      "14156",
      "17511",
      "18102",
```



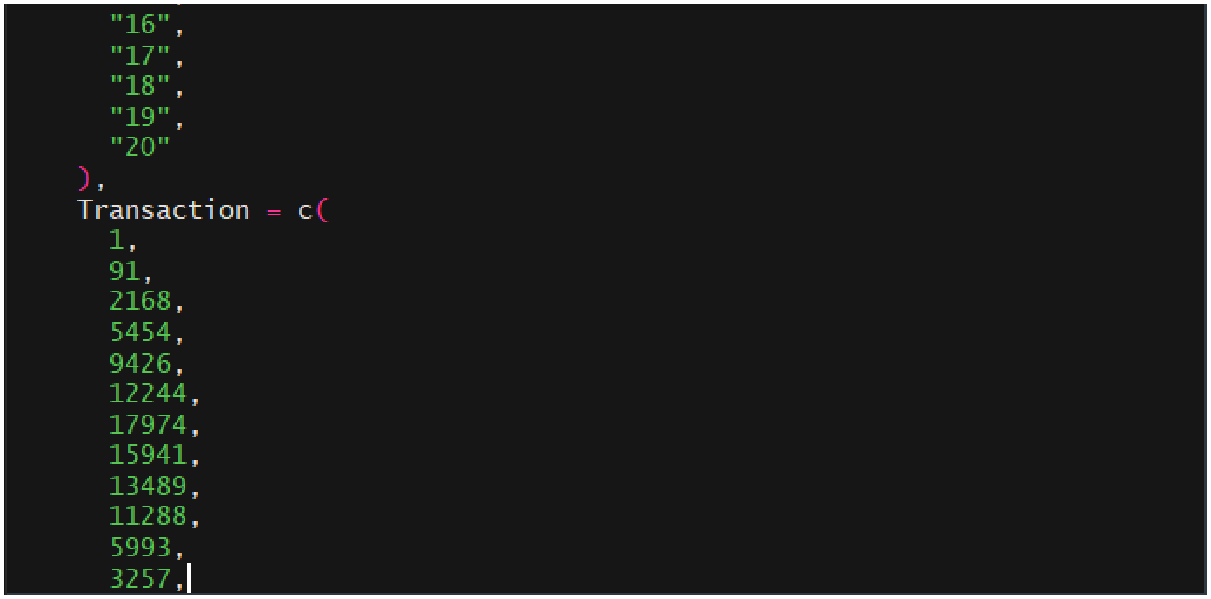
```
    "14298"
  ),
  Quantity = c(
    80995,
    48438,
    20831,
    19962,
    18686,
    18278,
    15806,
    15725,
    15514,
    12860
  )
)
ggplot(topcust, aes(x = reorder(CustomerId,-Quantity), y = Quantity)) +
geom_bar(stat = "identity", col = "forestgreen", fill = "darkorange1") +
labs(x = "Customer Id",
     y = "Quantity of Product Purchased",
```



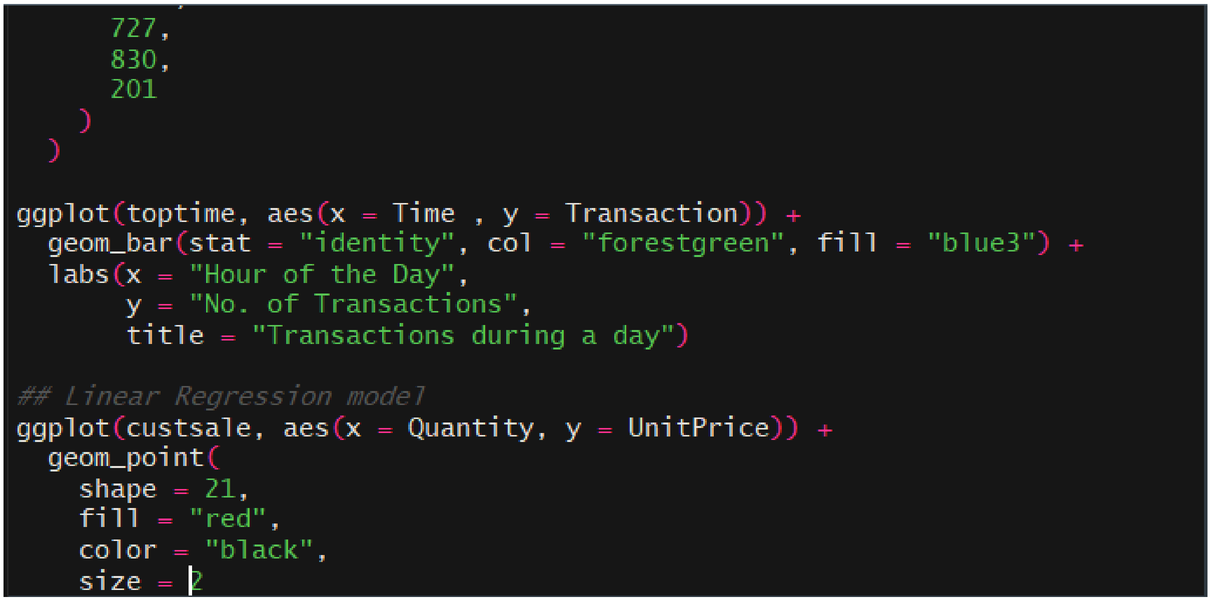
```
       title = "Top Buying Customers")

##Week day wise sales
weeksales <-
  data.frame(
    Week = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"),
    Revenue = c(
      199442.951,
      329886.41,
      402489.461,
      397773.65,
      485039.73,
      478045.15,
      0
    ),
    order = seq(1:7)
  )
ggplot(weeksales, aes(reorder(Week, order), Revenue)) +
  geom_bar(stat = "identity", col = "forestgreen", fill = "purple") +
```



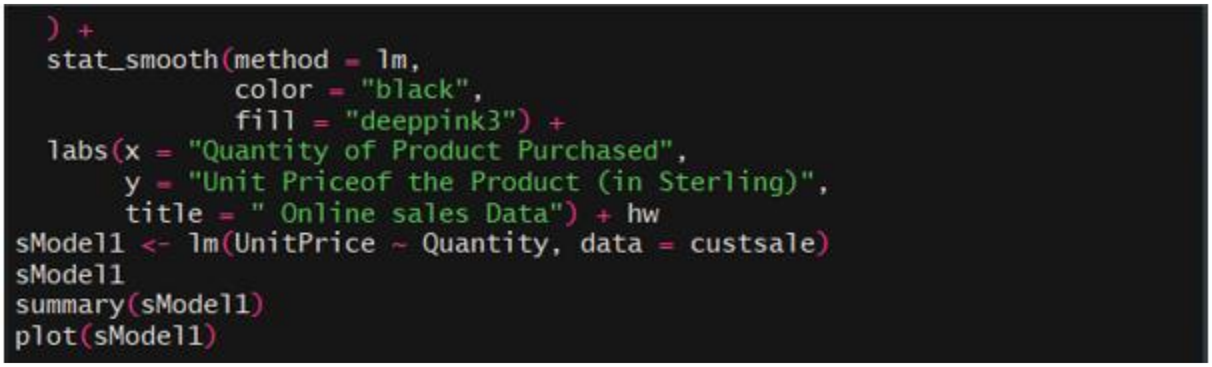
```
  labs(x = "Day of the Week",
       y = "Revenue",
       title = "Day wise Total Sales")


## Hourly Sales
toptime <-
  data.frame(
    Time = c(
      "6",
      "7",
      "8",
      "9",
      "10",
      "11",
      "12",
      "13",
      "14",
      "15",
```



```
      "16",
      "17",
      "18",
      "19",
      "20"
    ),
    Transaction = c(
      1,
      91,
      2168,
      5454,
      9426,
      12244,
      17974,
      15941,
      13489,
      11288,
      5993,
      3257,
```



```
      727,
      830,
      201
    )
  )

ggplot(toptime, aes(x = Time , y = Transaction)) +
  geom_bar(stat = "identity", col = "forestgreen", fill = "blue3") +
  labs(x = "Hour of the Day",
       y = "No. of Transactions",
       title = "Transactions during a day")

## Linear Regression model
ggplot(custsale, aes(x = Quantity, y = UnitPrice)) +
  geom_point(
    shape = 21,
    fill = "red",
    color = "black",
    size = 2
```



```
  ) +
  stat_smooth(method = lm,
              color = "black",
              fill = "deeppink3") +
  labs(x = "Quantity of Product Purchased",
       y = "Unit Priceof the Product (in Sterling)",
       title = " Online sales Data") + hw
sModel1 <- lm(UnitPrice ~ Quantity, data = custsale)
sModel1
summary(sModel1)
plot(sModel1)
```