2.0 Exploratory Data Analysis

**3.1 Data Inspection**

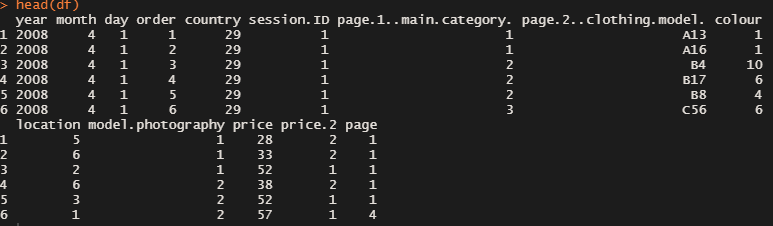


Figure 1.0

The figure above visualises the initial 6 rows of the dataset. The data comprises of 14 columns which is explained with the table below:

|  |  |
| --- | --- |
| **Columns** | **Explanation** |
| year | Year is 2008 |
| month | From April (4) to August (8) |
| day | Day number of the month |
| order | The sequence of clicks in one session |
| country | Indicating the country by identifying the origin of the IP Address |
| session.ID | ID of the session |
| page.1..main.category. | Main Product Category (1 to 4) |
| page.2..clothing.model. | Code for each product |
| colour | Colour of the product |
| location | Photo location on the page |
| model.photography | Model of the photo – enface or profile |
| price | Price in US Dollar |
| price.2 *(Target Variable)* | Variable informing whether price of the product is higher than the average price for the entire product category |
| page | Page number within the e-store website |

Table 1.0

The “price.2” is the target variable for this study that informs us whether price of the particular product is higher (1-Yes) or lower (2-No) than the average price for its entire product category.

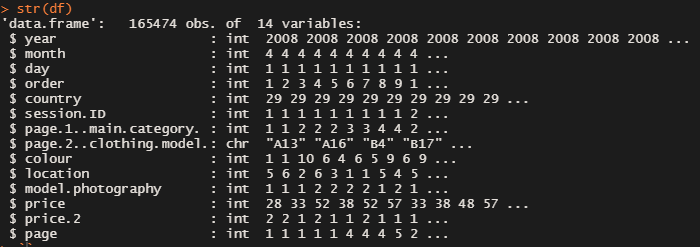


Figure 1.1

Except for the column “page.2..clothing.model.” which has the class of characters, all of the columns in the dataset have the class of integers. All of the columns except “price” were converted to data class “factor” for categorizing purposes and the data column “price” was converted into class of “numerical” as it is a continuous data.

****

Figure 1.2

The target variable “price.2” will be a “factor” with levels “1” and “2” after the class conversion.

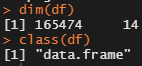


Figure 1.3

The dataset is stored in a data frame with a dimension of 165,474 rows with 14 columns which is the variables of the dataset.

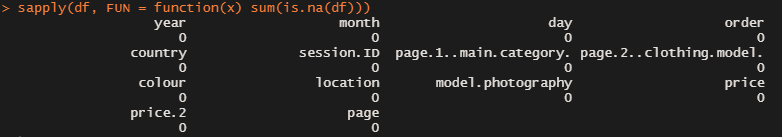


Figure 1.4

As we can see, there is no missing values from each of the columns from clickstream data.

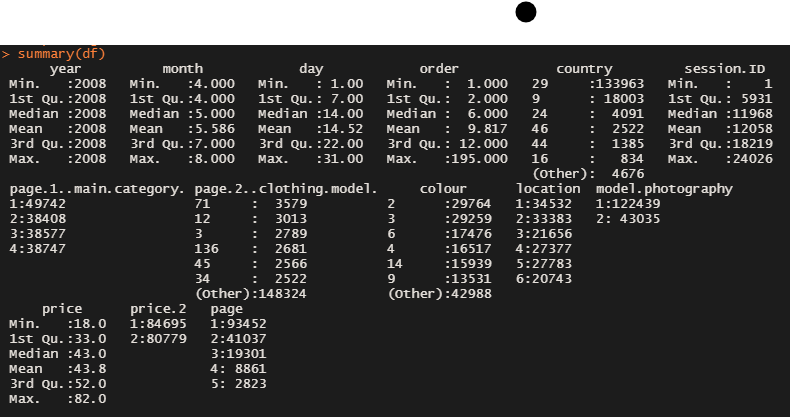


Figure 1.5.1

As we’re going through the summary of the dataset after the class conversion, we’re able to see the values in ascending order.

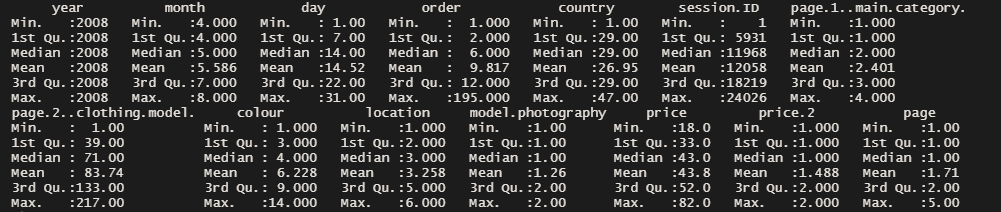
****

Figure 1.5.2

From the summary of the dataset in numeric class, we can see that some variables like model.photography, and price.2 which are one-hot encoding data. The variable “price” has quite a large variation and would be normalise if needed to improve the predictive models we build. The figures below shows the distribution of the data in plots deemed needed.

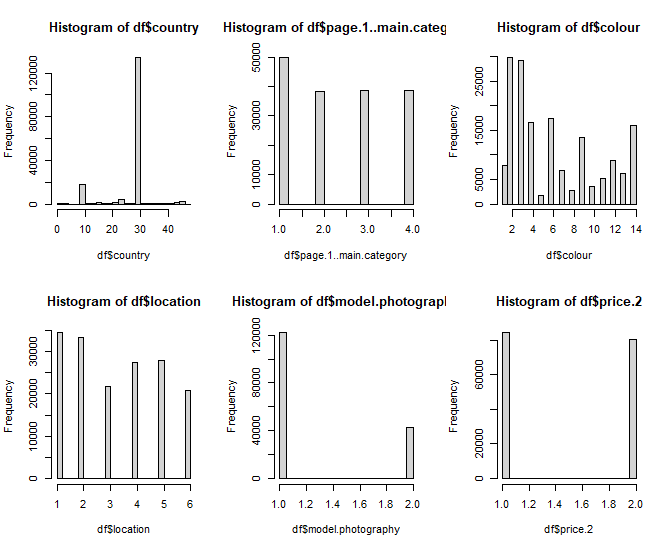


Figure 1.6.1

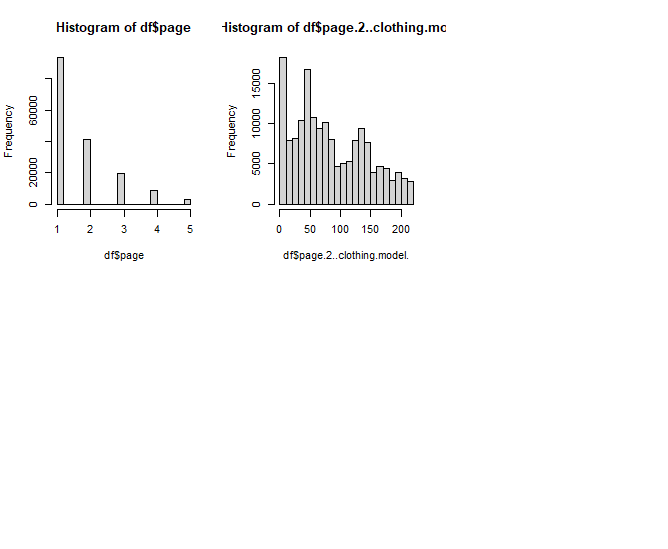


Figure 1.6.2

Most of the data collected was from Poland (29) and mostly sales were made in the first page of the e-shop. Our target variable “price.2” has a ratio of 84695:80779 which is considered balanced.

**3.2 Data Analysis**

We plotted several graphs to visualise and understand about the distribution of total sales, number of pieces sold, and its price.2.

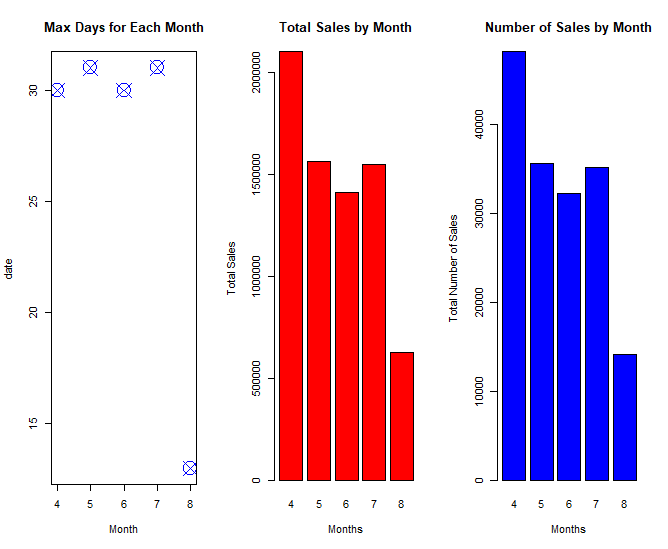


Figure 1.6.3

From figure 1.6.3, we can see that the data collected in August 2019 was only until day 13. This contributed to lower sales volume in August too. Moreover, April was the most profitable month in terms of clothes sold and sales.

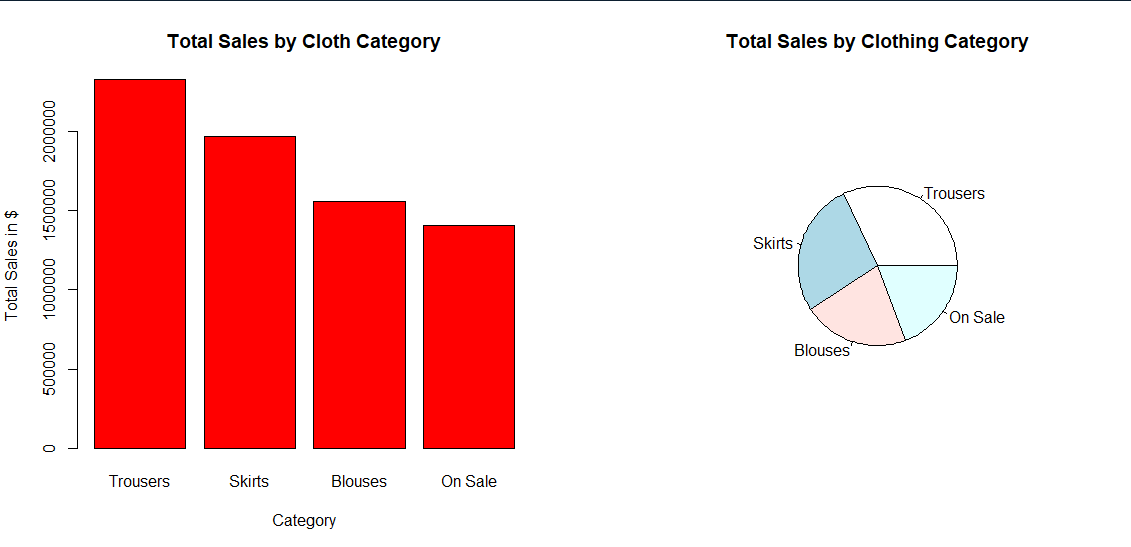
****

Figure 1.6.4

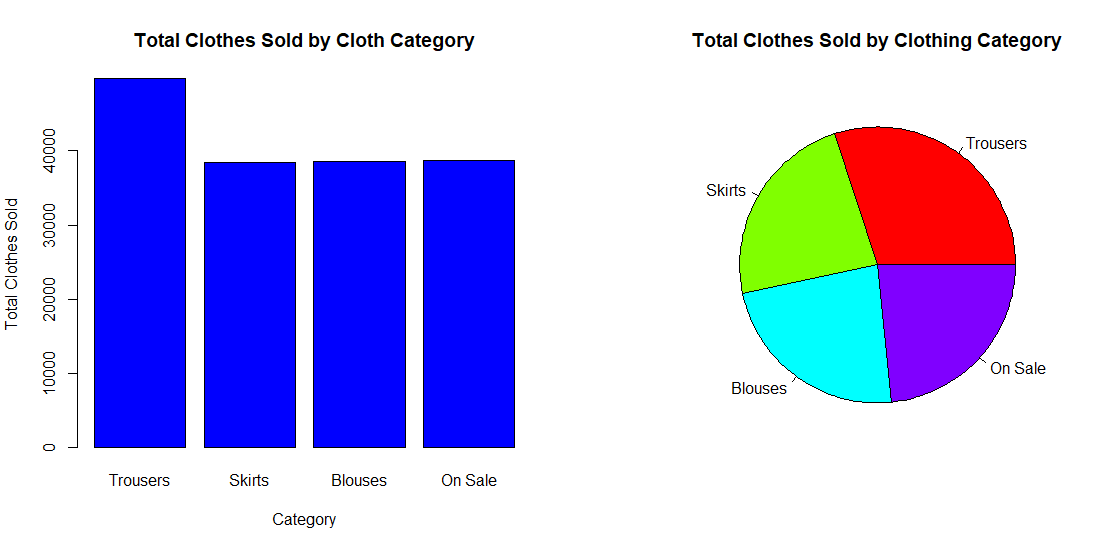


Figure 1.6.5

From figure 1.6.4 and 1.6.5, Trousers had the highest number of sales and number of pieces sold. However, clothes on sale was the worst performing category in these two measurements. It is also worth mentioning that Skirts was the second highest revenue even though it has roughly the same number of pieces sold with blouses and clothes on sale.

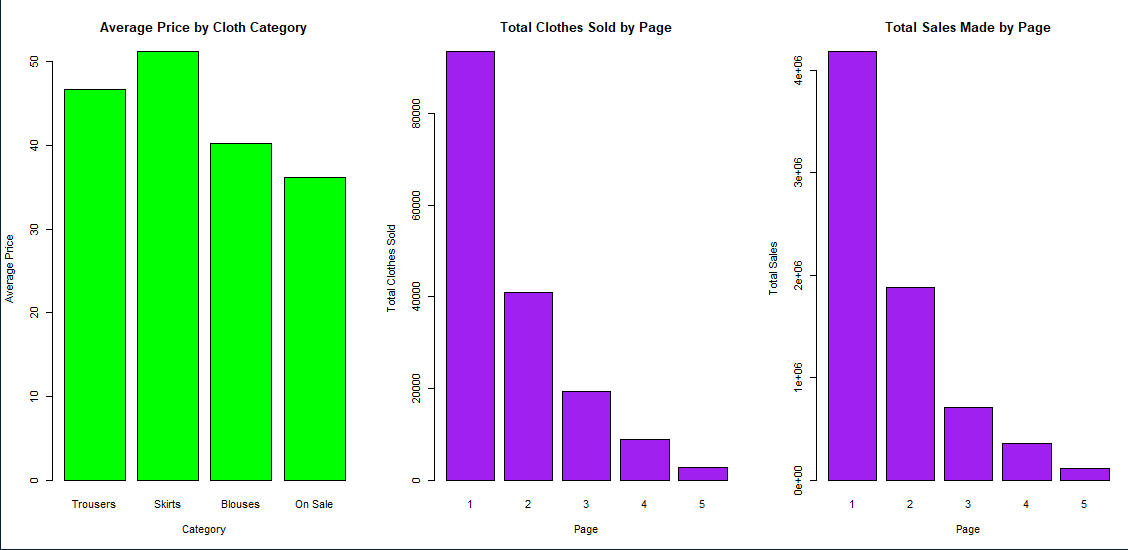


Figure 1.6.6

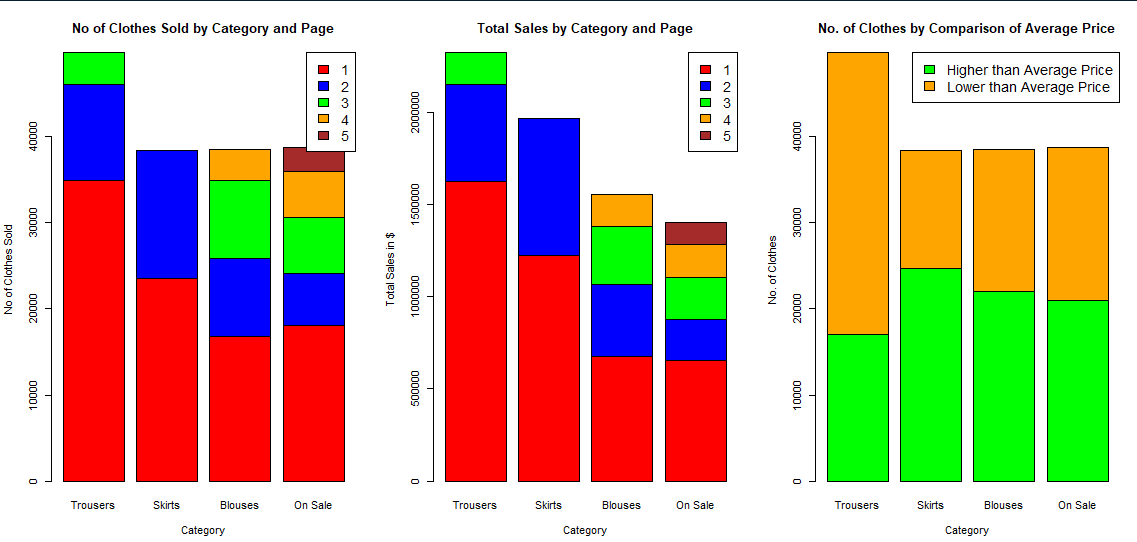


Figure 1.6.7

Some trends were noticeable from these charts in Figure 1.6.6 and 1.6.7:

* The average price for skirts was the highest among the clothing category.
* Many trousers were sold at lower than average prices whereas skirts were sold mostly at higher than average price
* Sales would decrease from page 1 to page 5.
* Trousers got the highest number of pieces sold in Page 1
* Skirts were sold quite well in page 2.
* Clothes on sale was the only category to be sold on page 1 to page 5

**3.3 Descriptive Statistics**

To understand how the distribution of the dataset looks like, we check the skewness and kurtosis of the dataset. Skewness is the measurement of symmetry, whereas kurtosis is the measurement of the weight of the tails relative to a normal distribution. We checked the skewness and kurtosis of the following variables:



Figure 1.7



Figure 1.8

As a rule of thumb for skewness a distribution is skewed to the right if the value is positively skewed and vice versa. The variable “country” is highly skewed to the left, whereas “model.photography” and “page” are highly skewed to the right. The variable “colour” is moderately skewed to the right and the rest of the columns are approximately skewed to the right.

Furthermore, a normally distributed graph would have a kurtosis of 3. Kurtosis of higher than 3 would have longer and fatter tails, whereas a lower than 3 would have thinner and shorter tails. The only 2 variables that have longer and fatter tails are “country” and “page”, while the rest had thinner and shorter tails.

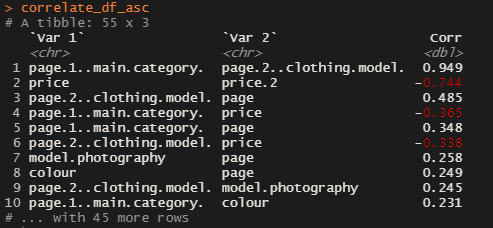


Figure 1.9.1

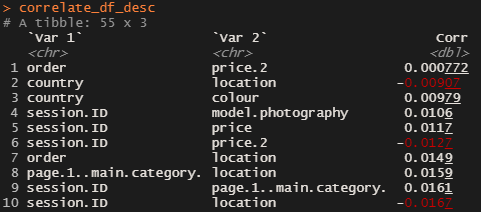


Figure 1.9.2

Correlation coefficient measures the linear relationship between 2 variables. A positive (direct) linear relationship would have a value of 0 to 1, whereas a negative (inverse) linear relationship would have values from -1 to 0. There would be no relationship between variables when the value is 0.

By tabulating a list of correlation coefficient between all the variables, we take the 10 highest and lowest correlation pairs of variables including the absolute values. We can see that there is a strong positive correlation between “page.1..main.category.” and “page.2..clothing.model.” and a strong negative correlation between “price” and “price.2” as both values are more than 0.50. The rest of the variable pairs from figure 1.9.1 are pairs with some relationship.