Analysis on Car Sales using SQL, R and Tableau

(Submitted by: Prathigadapa Rashmi, M12853369)

Data Description:

The dataset contains monthly sales of passenger cars in Norway by make and model of the cars. I have taken this dataset from Kaggle. You can download the data from below URL: https://www.kaggle.com/dmi3kno/newcarsalesnorway#norway_new_car_sales_by_make.csv

This is a CSV file and it contains 6 data columns namely, 'Year', 'Month', 'Make', 'Model', 'Quantity' and 'Percentage share'. It has 5M lines, corresponding to car sales for a decade (from 2007 to 2017). Below is the sample data of the first few rows.

	Year	Month	Make	Model	Quantity	Pct
1	2007	1	Volkswagen	Volkswagen Passat	1267	10.0
2	2007	1	Toyota	Toyota Rav4	819	6.5
3	2007	1	Toyota	Toyota Avensis	787	6.2
4	2007	1	Volkswagen	Volkswagen Golf	720	5.7
5	2007	1	Toyota	Toyota Corolla	691	5.4
6	2007	1	Peugeot	Peugeot 307	481	3.8

Year: This column denotes the Year in which the car is sold. It has values from 2016, January to 2017 January.

Month: This column denotes the Month in which the car is sold.

Make: This column denotes the Manufacturer brand of the car. Ex: Volkswagen, Toyota etc.

Model: This column denotes the Model of the car. Ex: BMW-i3, Volkswagen Golf, Tesla S75

Quantity: This column shows the number of cars sold.

Pct: This column denotes the percent share in monthly total.

Data Normalization:

Data Normalization is decomposing tables to eliminate data redundancy and undesirable characteristics. In our data, we can perform normalization to segregate the make and model of the car into a different table. There are 23 values within the Make column which can be

mapped to multiple rows. Similarly, 89 unique values of model can be segregated into a different dataset. So that it is efficient and convenient to handle the data flow.

Data Preparation:

To make sure that the data is clean, we need to perform few checks.

- Check for duplicate values: We do not have any duplicates in our data.
- Change bad header names to readable format. The name of the column 'Pct' is not understandable and therefore can be changed to 'PercentShare'
- Added a unique column named 'CarID' to identify the data at unique level.

Data Statistics using SQL:

We can perform basic statistics using SQL and R. Below are the list of statistics obtained from the data.

- Total number of rows in the data is 2694 lines.
- Top 3 Cars based on the popularity (Quantity sold). Below is the list of top three cars:

	Make	Model	Quantity
1	Volkswagen	Volkswagen Golf	85787
2	Volkswagen	Volkswagen Passat	40575
3	Toyota	Toyota Auris	36668

- Top 3 cars in percent of shares is same as the top 3 popular cars.
- The year with maximum number of sales is '2016', with 93,601 cars sold in that year. In that year, maximum sales are in the month of April.
- Maximum, Minimum and Average sales by Make.

	make	Maximum_sal...	Minimum_sal...	Average_sales
1	Volkswagen	1713	55	387.925
2	Tesla	1493	36	306.324324324324
3	Mitsubishi	739	70	277.314285714286
4	Toyota	819	25	273.691056910569
5	Nissan	716	13	264.338888888889

- Brand that produced highest number of cars.

	Count	Make
1	170687	Volkswagen
2	134656	Toyota
3	71558	Volvo
4	57763	Ford
5	47581	Nissan

- Brand that produced least number of cars.

	Count	Make
1	129	Citroen
2	184	Mercedes-Benz
3	1025	Saab
4	1713	Subaru
5	2254	Hyundai

- Overall, the month with maximum number of sales is 'January' with total number of 66,091 cars being sold in the last decade.
- Least sold cars in the last decade. Below is the list:

	Make	Model	Tot_Quan...
1	Mercedes-Benz	Mercedes-Benz CLA	61
2	Skoda	Skoda Rapid	63
3	Mercedes-Benz	Mercedes-Benz GLK	64

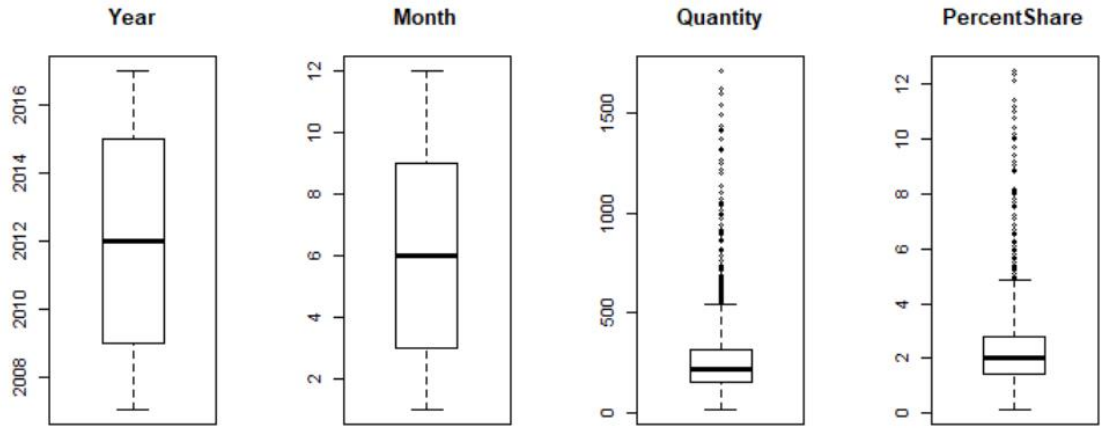
- Top 5 car count based on model, make and year.

	year	make	model	count_...
1	2007	Audi	Audi A4	12
2	2007	Ford	Ford Focus	12
3	2007	Honda	Honda CR-V	12
4	2007	Opel	Opel Astra	12
5	2007	Toyota	Toyota Avensis	12

Data Analysis using R:

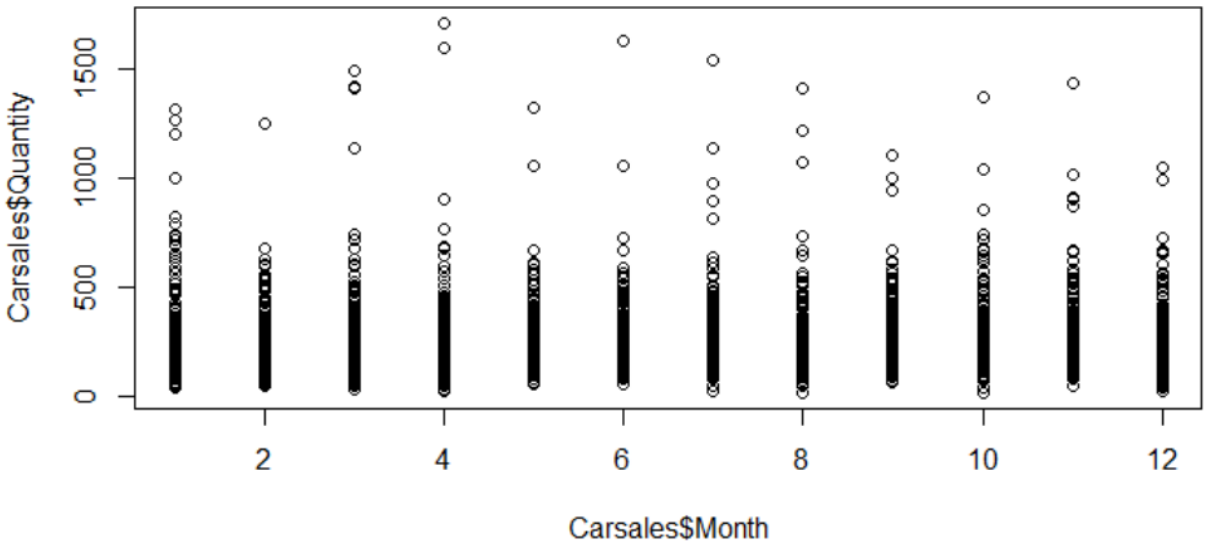
We can perform below statistics using R:

- To identify outliers in our data, we can draw box plots for all the quantitative fields. We see that the fields Quantity and PercentShare have outliers because they vary based on the need and it is acceptable to have outliers in such columns.



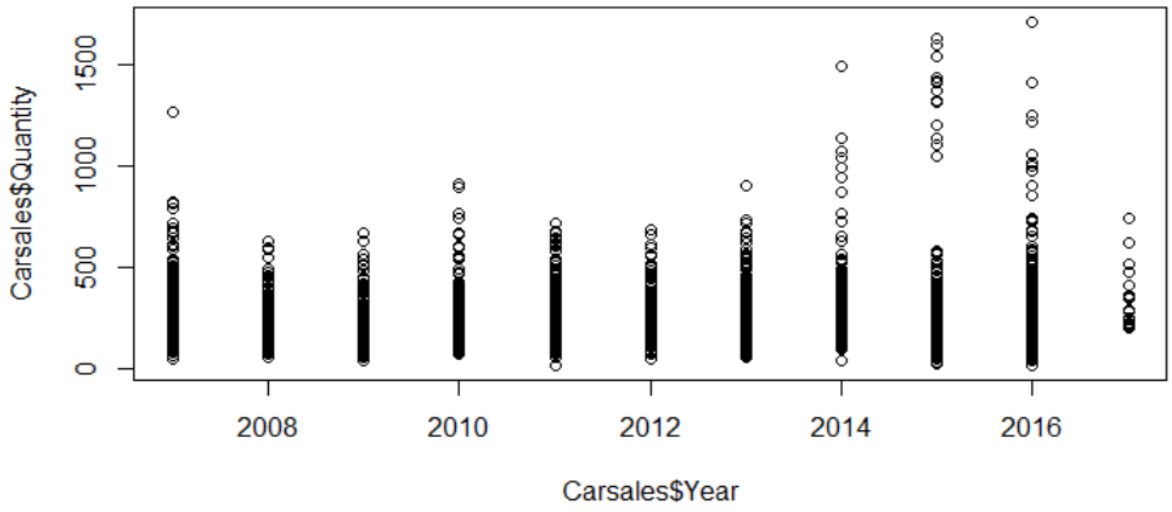
- To identify the distribution of cars across each month.

Month vs Quantity



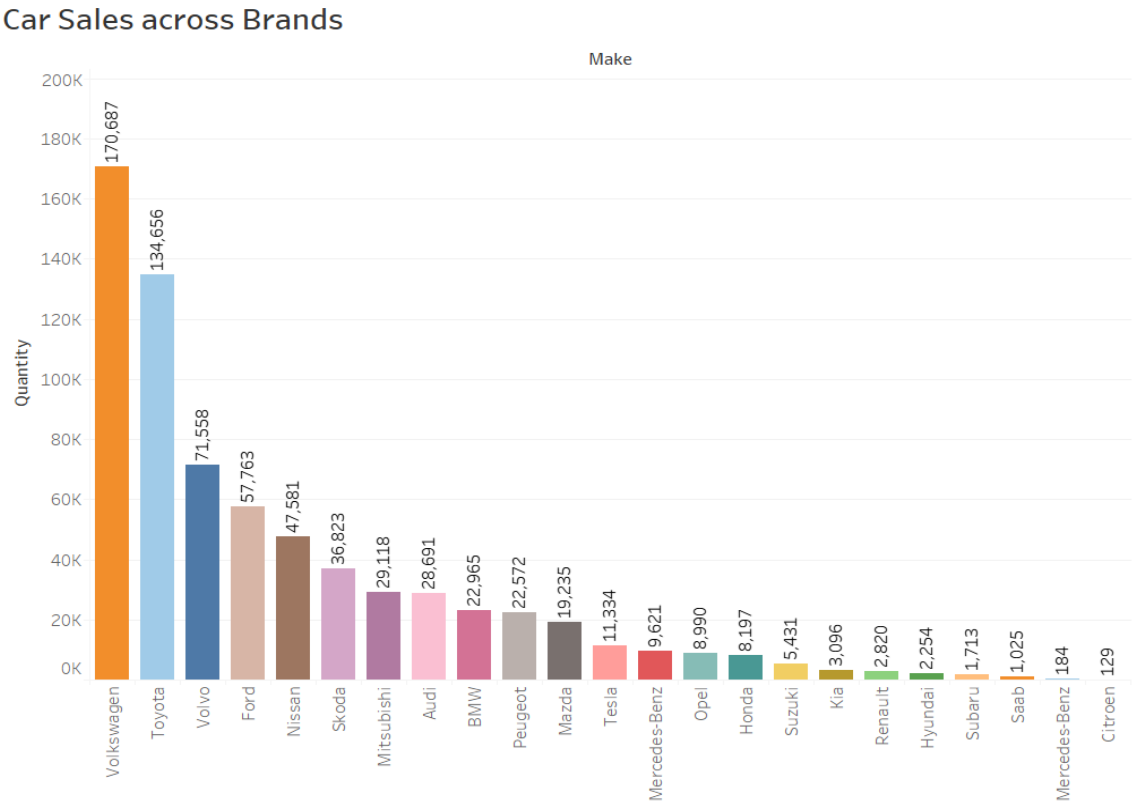
- To identify the trend of cars sold over the decade.

Year vs Quantity

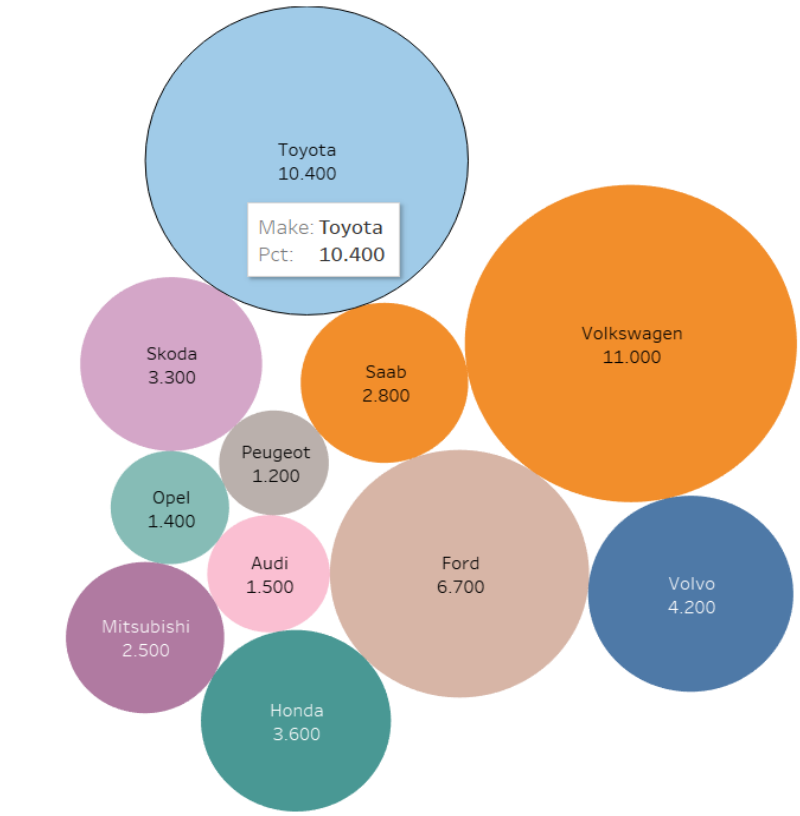


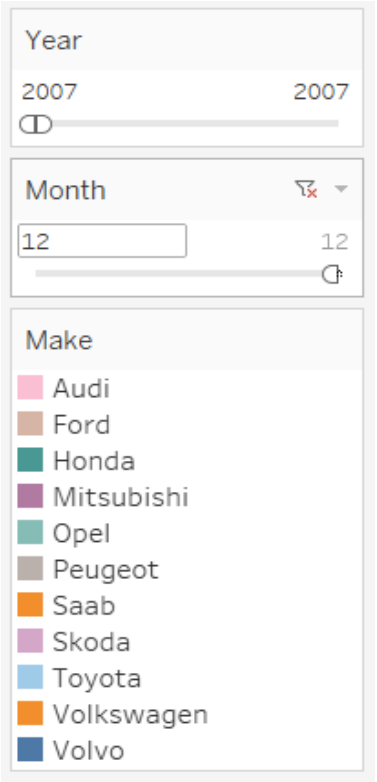
Data Analysis using Tableau:

Below are the visualizations obtained using Tableau.



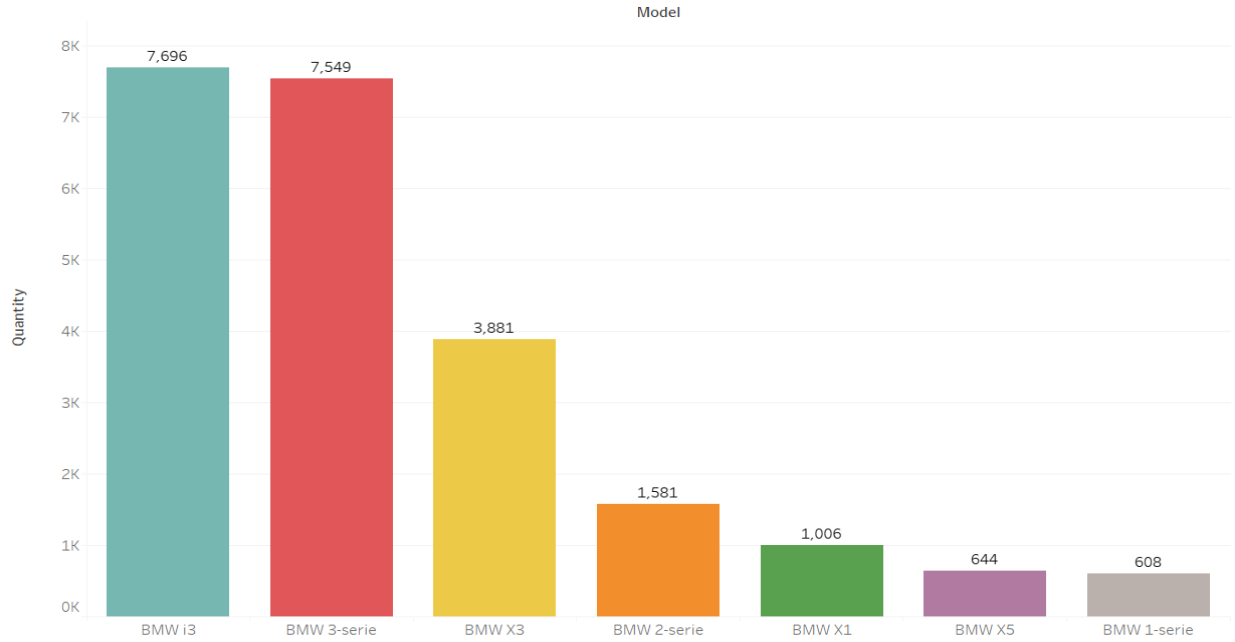
Yearly and Monthly Percent Shares





(Filters to set year and month values)

Model-wise Sales Analysis



(In the above chart, we can filter for multiple make and models)

Key Findings:

- **From the above analysis, it can be said that, the most popular cars in Norway are Volkswagen Golf and Passat. Second popular brand is Toyota.**
- **The top sold cars are average priced passenger cars like Volkswagen, Toyota, Nissan and Ford.**
- **The manufacturers like Audi, BMW, Peugeot are medium range cars sold. Even though they are bit expensive, people prefer them may be because of the make and various models available.**
- **The costliest brands like Tesla, Mercedes Benz, BMW and sports cars like Subaru are least sold cars because of the high prices and may be because similar features are being offered by medium range cars. So, they are not willing to spend a lot on these cars.**