Instituto de Educação Superior de Brasília - IESB
Ciência de Dados e Inteligência Artificial

Análise exploratória sobre venda de carros na Noruega

por

Victor Augusto Souza Resende

Brasília - DF, 8 de Agosto de 2020

Vendas de Carros na Noruega

Trabalho para aplicação de técnicas estudadas no segundo semestre do curso de Ciência de Dados & Inteligência Artificial na matéria **Estatística**, juntamente com estudos particulares do autor sobre séries temporais e modelos simples de previsões ensinados no Curso de Inverno na matéria **R na prática**, ambos cursos citados ministrados pelo Instituto de Educação Superior de Brasília.

Dedicatória

À minha família, em especial aos meus pais, minha namorada e meus amigos, pessoas da qual obtive muito apoio.

Agradecimentos

Primeiramente, agradeço ao meu primo Marcelo e ao meu professor Suélio por tornar a documentação desse trabalho viável via LaTeX. Agradeço também à minha família e minha namorada Stefany pelo apoio e aos meus professores por grandes ensinamentos.

RESUMO

RESENDE, V.A. **Analise exploratória sobre venda de carros na Noruega nos anos de 2007 até o fim de 2016**. 2020. Ciência de Dados Inteligência Artificial, Instituto de Educação Superior de Brasília, Brasília, 2020.

A Noruega por ser um país com grande facilidade em questões burocráticas tornou viável a coleta de dados apresentado nesse trabalho, e visando aplicar ensinamentos conquistados cursando até então o segundo semestre de Ciência de Dados Inteligência Artificial, foi efetuada uma análise exploratória sobre tal tema para encontrar pontos específicos e intuições de negócios com o uso de técnicas estatísticas como séries temporais e previsões com uso das ferramentas R e RStudio. Toda a base de dados, intitulada como "New Car Sales Norway", se encontra disponível na plataforma Kaggle. Esse trabalho e o código utilizado se encontram no repositório do author na plataforma GitHub: `github.com/victoresende19`.

Palavras-chave: Análise Exploratória de Dados, R, RStudio, Venda de Carros, Noruega.

Abstract

RESENDE, V.A. **Exploratory analysis on car sales in Norway from 2007 to the end of 2016**. 2020. Data Science Artificial Intelligence, Brasília Institute of Higher Education, Brasília, 2020.

Norway, being a country with great ease in bureaucratic issues, made the data collection presented in this work viable, and in order to apply lessons learned while studying the third semester of Data Science and Intelligence Artificial, an exploratory analysis was carried out on this topic to find specific points and business sights using statistical techniques such as time series and forecasts using the R and RStudio tools. The entire database, entitled "New Car Sales Norway", is available on the Kaggle platform. This document and the scrip used were searched in the author's repository on the GitHub platform: `github.com/victoresende19`.

Key-words: Exploratoty Analysis, R, RStudio, Car Sales, Norway.

Sumário

1 Conhecendo a Base de Dados **3**
1.1 Introdução . . . 3
1.2 Dicionário da Base de Dados . . . 4
1.2.1 Notas sobre as variáveis . . . 5
1.2.2 Notas sobre a base de dados . . . 5

2 Limpeza dos dados **6**
2.1 Quantidade de Dados faltantes . . . 6
2.2 Excluindo Dados faltantes . . . 7

3 Exploração de Dados - Estatísticas Descritivas **8**
3.1 Medidas de tendência central . . . 8
3.1.1 Medidas de tendência central - Quantidade . . . 8
3.1.2 Medidas de tendência central - Quantidade & Marca . . . 9
3.1.3 Medidas de tendência central - Importações . . . 10
3.1.4 Box Plot . . . 11
3.2 Medidas de dispersão . . . 13
3.2.1 Medidas de dispersão - Quantidade . . . 14
3.2.2 Medidas de dispersão - Quantidade & Marca . . . 14
3.2.3 Medidas de dispersão - Importações . . . 15
3.3 Formato da distribuição dos dados . . . 16
3.3.1 Formato da distribuição dos dados - Quantidade . . . 17
3.3.2 Formato da distribuição dos dados - Quantidade & Marca . . . 18
3.3.3 Formato da distribuição dos dados - Importações . . . 20
3.4 Correlação entre as variáveis quantitativas . . . 21
3.4.1 Ideia geral - Exemplo: Quantidade & Quantidade Diesel . . . 21
3.4.2 Correlação das variáveis - Geral . . . 23

4 Análise das medidas encontradas **26**
4.1 Introdução às análises . . . 26
4.1.1 Análise - Quantidade . . . 26
4.1.2 Análise - Quantidade & Marca . . . 27
4.1.3 Análise - Importação . . . 28

5 Regressões e Previsões **29**
5.1 Regressão Linear . . . 29
5.2 Regressão Linear - CO2 vendidos vs. Ano . . . 30
5.3 Regressão Linear - Quantidade vendidos vs. Ano . . . 31
5.3.1 Série Temporais . . . 32

5.3.2 Série Temporal - Quantidade vs. Ano 32
5.3.3 Série Temporal - Quantidade vs. Mês e Ano 33
5.3.4 Previsão 34
5.3.5 Provando a previsão encontrada 35

6 Conclusão **36**

Capítulo 1

Conhecendo a Base de Dados

1.1 Introdução

O setor automotivo de um país é seguramente um dos setores econômicos que certamente pode-se fazer diversas análises a serem refletidas na população. Por exemplo, se um país vende diversos veículos, isso pode dizer que, de certa maneira, a economia do país esta crescendo frequentemente pois a população possui renda para efetuar a compra de tais veículos. Porém por outro lado, se um país tem queda no setor automotivo, podemos ter algumas alternativas nebulosas para se entender. Por exemplo, a população pode estar empobrecida e não possui dinheiro para efetuar a compra de veículos automotivos, o país pode estar enfrentando alguma crise econômica (Seja interna ou global), ou então, a população está escolhendo um caminho mais saudável em relação à natureza, como por exemplo, os cidadãos podem estar optando por bicicletas, se locomover por meio de transporte público ou alguma outra alternativa que seja menos poluidora.

Nesse trabalho será analisado, por meio de métodos estatísticos e com auxílio de ferramentas de software, a venda de carros na Noruega, importação de veículos efetuado pelo país no período de ano 2007 até o fim de 2016. Será analisado também especificamente algumas marcas de carros escolhidas aleatoriamente pelo autor usando artifícios como por exemplo, visualização de correlação entre variáveis, Regressões Lineares em relações às certas variáveis, séries temporais e modelos básicos de previsões. Vale ressaltar que todas as aplicações passaram por estudo prévio e, posteriormente, implementação usando as ferramentas R e RStudio e possui referências obtidas na leitura dos livros: Estatística Básica por Morettin e BUSSAB (2017), The R book por Crawley (2012) e Data science for business por Provost e Fawcett (2013).

Vale ressaltar que essa base de dados vai de janeiro de 2007 até janeiro de 2017, porém para trabalharmos com analises certeiras para cada ano, a data de janeiro de 2017 não será contabilizada nos cálculos. A data excluída servirá apenas posteriormente para provarmos o modelo de previsão. Este estudo foi concluído em apenas um mês apenas com o objetivo de aplicação de técnicas aprendidas, sendo assim o tempo de confecção foi levado muito em consideração na escolha das variáveis para que as análises pudessem ser o mais eficaz possível.

1.2 Dicionário da Base de Dados

A base de dados está disponível em:

Kaggle - New Cars Sales Norway.

Como todas as variáveis estão em inglês, aqui será explicado o significado de cada

- **Year:** Ano de venda.
- **Month:** Mês de venda.
- **Quantity:** Quantidade vendida de automóveis (Geral).
- **Quantity YoY:** Quantidade vendida de automóveis Ano a Ano.
- **Import:** Quantidade de importações de automóveis (Automóveis usados).
- **Import YoY:** Quantidade de importações de automóveis Ano a Ano.
- **Used:** Quantidade vendida de automóveis usados.
- **Used YoY:** Quantidade vendida de automóveis usados Ano a Ano.
- **Avg CO2:** Emissão média de CO2 de automóveis vendidos.
- **Bensin CO2:** Emissão média de CO2 de abastecimento com Bensina.
- **Diesel CO2:** Emissão média de CO2 de abastecimento com Diesel.
- **Quantity Diesel:** Quantidade de carros abastecidos com Diesel vendidos.
- **Diesel Share:** Quantidade de automóveis a Diesel vendidos.
- **Diesel Share Ly:** Quantidade de automóveis a Diesel vendidos há um ano.
- **Quantity Hybrid:** Quantidade de automóveis híbridos novos vendidos.
- **Quantity Eletric:** Quantidade de automóveis elétricos novos vendidos.
- **Import Eletric:** Quantidade de importações de automóveis elétricos usados.
- **Make:** Marca do automóvel vendido.
- **Model:** Modelo do automóvel vendido.

1.2.1 Notas sobre as variáveis

As seguintes variáveis: *Quantity Hybrid* e *Quantity Eletric* apenas começaram a serem postados a partir do ano 2011.

Em relação à variável *Diesel Share* foi utilizada a partir da seguinte formúla: $\frac{QuantityDiesel}{Quantity}$.

Como pôde ser evidenciado, a base de dados possui algumas variáveis com determinada denominação *YoY*. YoY é uma abreviatura para *Year Over Year*, ao pé da letra poderia ser traduzida para a língua portuguesa como Ano a Ano. YoY é uma medida usada para comparar crescimento de determinado seguimento atual em comparação ao mesmo seguimento num período passado, podendo assim calcular a taxa de crescimento ou decrescimento do seguimento analisado. Vale lembrar que tal medida é comumente retornada em porcentagem. Sendo assim temos que a formula de YoY é dada por:

$$\frac{PerodoAtual - PerodoPassado}{PerodoPassado} * 100.$$

É importante citar que nesse trabalho será avaliado de maneira mais específica as seguintes variáveis, pois este trabalho tem como objetivo apenas provar estudos realizados:

- Quantity
- Quantity & Make
- Import

Tal escolha se deu por uma melhor associação das variáveis com as análises que serão feitas juntamente com a ideia de serem efetuadas análises rasas porém eficientes, já que demandaria muito tempo efetuar as mesmas com a vasta gama de variáveis presentes no banco de dados utilizado.

1.2.2 Notas sobre a base de dados

A base de dados é apresentada com todas as variáveis citadas anteriormente em relação à cada mês de acordo com determinado ano, como será apresentado na seguinte imagem:

	Year	Month	Quantity	Quantity_YoY	Import	Import_YoY	Used	Used_YoY	Avg_CO2	Bensin_Co2
1	2007	1	12685	5227	2276	257	NA	NA	152	155
2	2007	2	9793	2448	1992	-89	NA	NA	156	159
3	2007	3	11264	1445	2626	45	NA	NA	159	161
4	2007	4	8854	504	2220	-130	NA	NA	160	165
5	2007	5	12007	1592	2881	7	NA	NA	160	163
6	2007	6	11083	1545	3038	23	NA	NA	161	163
7	2007	7	12062	1908	3768	137	NA	NA	159	161
8	2007	8	10786	1993	3419	260	NA	NA	160	160
9	2007	9	9340	498	2897	-28	NA	NA	160	160
10	2007	10	11646	2973	3185	597	NA	NA	159	160
11	2007	11	10453	1709	2957	544	NA	NA	160	161
12	2007	12	9222	-1811	2097	-1257	NA	NA	162	162
13	2008	1	9901	-2784	2287	11	NA	NA	158	155
14	2008	2	10567	774	2627	635	NA	NA	160	159
15	2008	3	9506	-1758	2270	-356	NA	NA	159	160
16	2008	4	11704	2850	2930	710	NA	NA	159	160

Figura 1.1: Exemplo base de dados

Capítulo 2

Limpeza dos dados

2.1 Quantidade de Dados faltantes

Após conhecer as variáveis que serão trabalhadas, foi efetuada a seguinte função na ferramenta RStudio para descobrir a porcentagem de dados faltantes em cada variável:

```
NAS <- round(colSums(is.na(norway_car_month))*100/nrow(norway_car_month),2)
```

Sendo assim, apenas foi necessário chamar a nova variável (NAS) da seguinte forma:

$$\text{NAS} > 0$$

Para descobrir quais variáveis possuiam dados faltantes maior que zero. Sendo assim temos o seguinte resultado:

- **Used:** 49.59% de dados faltantes.
- **Used YoY:** 59.50% de dados faltantes.
- **Quantity Hybrid:** 39.67% de dados faltantes.
- **Quantity Eletric:** 39.67% de dados faltantes.
- **Import Eletric:** 56.20% de dados faltantes.

2.2 Excluindo Dados faltantes

Como a ideia desenvolvida nesse trabalho não visa trabalhar em cima de variáveis como Quantidade de automóveis Híbridos ou Elétricos vendidos/importados, as seguintes colunas foram excluídas:

- Quantity Hybrid
- Quantity Eletric
- Import Eletric

Sobrando assim apenas as seguintes variáveis:

- Used
- Used YoY

Variáveis que se referem às quantidades de automóveis usados vendidos. Porém, como elas possuíam altas taxas de dados faltantes, principalmente a variável *Used*, foi decidido que tais colunas também seriam excluídas, pois usar técnicas de preenchimento poderia ser arriscado e comprometer o resultado final da análise.

A exclusão de cada variável citada anteriormente se deu com o seguinte código em R:

```
norway_car_month$Used_YoY <- NULL
norway_car_month$Used <- NULL
norway_car_month$Import_Electric <- NULL
norway_car_month$Quantity_Hybrid <- NULL
norway_car_month$Quantity_Electric <- NULL
```

Mesmo a Noruega, na data de confecção deste trabalho, tornando-se um país que vende mais carros elétricos do que tradicionais (De acordo com a líder da Associação Norueguesa de Veículos Elétricos, Christina Bu) não foi possível fazer tal análise de maneira eficiente pois a base de dados era referente até o ano de 2016, no qual foi evidenciado que ainda possuía diversos dados faltantes, e o foco deste trabalho não são variáveis para carros híbridos/elétricos. Então, após analisar caso a caso visando o resultado final do trabalho, todas as variáveis que possuíam dados faltantes foram excluídas sem danos significativos à ideia final.

Capítulo 3

Exploração de Dados - Estatísticas Descritivas

3.1 Medidas de tendência central

As medidas de tendência central são medidas que representam o centro ou o meio de uma base de dados, algumas dessas medidas são: Média, Moda e Mediana. Porém nesse trabalho não iremos trabalhar com a Moda pelo seguinte motivo:

- Para a variável Quantidade não faz sentido analisar a moda pois todo mês, durante os anos, teremos, certamente números de quantidade de vendas diferentes.
- Para a variável Quantidade e Modelo não faz sentido analisar a moda pois todo mês, durante os anos, teremos certamente números de quantidade de vendas diferentes.
- Para a variável Importação não faz sentido analisar a moda pois todo mês, durante os anos, teremos certamente números de importações diferentes.

3.1.1 Medidas de tendência central - Quantidade

A seguir será demonstrado o resultado das medidas de tendência centrais da variável ***Quantidade*** que se refere à quantidade de veículos vendidos nos período dos anos de 2007 até o fim de 2016:

- **Média:** A média da quantidade de veículos vendidos em todos os meses nos anos de 2007 até 2016 é aproximadamente 11134 veículos.

- **Mediana:** A mediana da quantidade de veículos vendidos em todos os meses nos anos de 2007 até 2016 é de 11385 veículos.

Medidas alcançadas facilmente por meio do seguinte código na linguagem R usando a ferramenta RStudio:

```
mean(norway_car_month$Quantity)
11134.3

median(norway_car_month$Quantity)
11385
```

3.1.2 Medidas de tendência central - Quantidade & Marca

A seguir será demonstrado o resultado das medidas de tendência centrais da variável ***Quantidade*** em relação à variável ***Marca*** visando a quantidade de veículos vendidos em relação às marcas de veículos *Ford* e *Toyota*.

As seguintes medidas de tendência central para a variável ***Quantidade*** e ***Ford*** são:

- **Média:** A média de quantidade de veículos da **Ford** vendidos em todos os meses nos anos de 2007 até 2016 é de 234 veículos.

- **Mediana:** A mediana de quantidade de veículos da **Ford** vendidos em todos os meses nos anos de 2007 até 2016 é de 212 veículos.

As seguintes medidas de tendência central para a variável ***Quantidade*** e ***Toyota*** são:

- **Média:** A média de quantidade de veículos da **Toyota** vendidos em todos os meses nos anos de 2007 até 2016 é de 273 veículos.

- **Mediana:** A mediana de quantidade de veículos da **Toyota** vendidos em todos os meses nos anos de 2007 até 2016 é de 268 veículos.

Medidas alcançadas facilmente por meio do seguinte código na linguagem R usando a ferramenta RStudio:

```
#FORD
mean(norway_car_model$Quantity[norway_car_model$Make == 'Ford'])
234.8089

median(norway_car_model$Quantity[norway_car_model$Make == 'Ford'])
212


#TOYOTA
mean(norway_car_model$Quantity[norway_car_model$Make == 'Toyota'])
273.6911

median(norway_car_model$Quantity[norway_car_model$Make == 'Toyota'])
268
```

3.1.3 Medidas de tendência central - Importações

A seguir será demonstrado o resultado das medidas de tendência centrais da variável ***Importação*** que se refere à quantidade de veículos importados usado no período do ano de 2007 até o fim de 2016.

- **Média:** A média da quantidade de veículos importados em todos os meses nos anos de 2007 até 2016 é aproximadamente 2204 veículos.

- **Mediana:** A mediana da quantidade de veículos importados em todos os meses nos anos de 2007 até 2016 é de 2263 veículos.

Medidas alcançadas facilmente por meio do seguinte código na linguagem R usando a ferramenta RStudio:

```
mean(norway_car_month$Import)
2204.372

median(norway_car_month$Import)
2263
```

3.1.4 Box Plot

A ferramenta estatística BoxPlot, presente também no software R e RStudio, pode nos trazer uma breve ilustração sobre medidas de tendência centrais, sendo representado e interpretado da seguinte maneira:

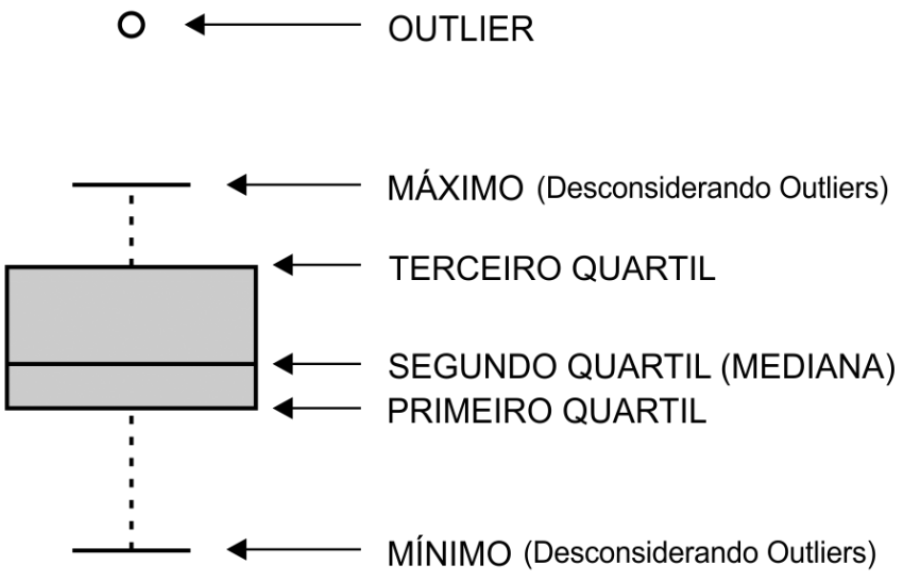


Figura 3.1: BoxPlot - Explicação

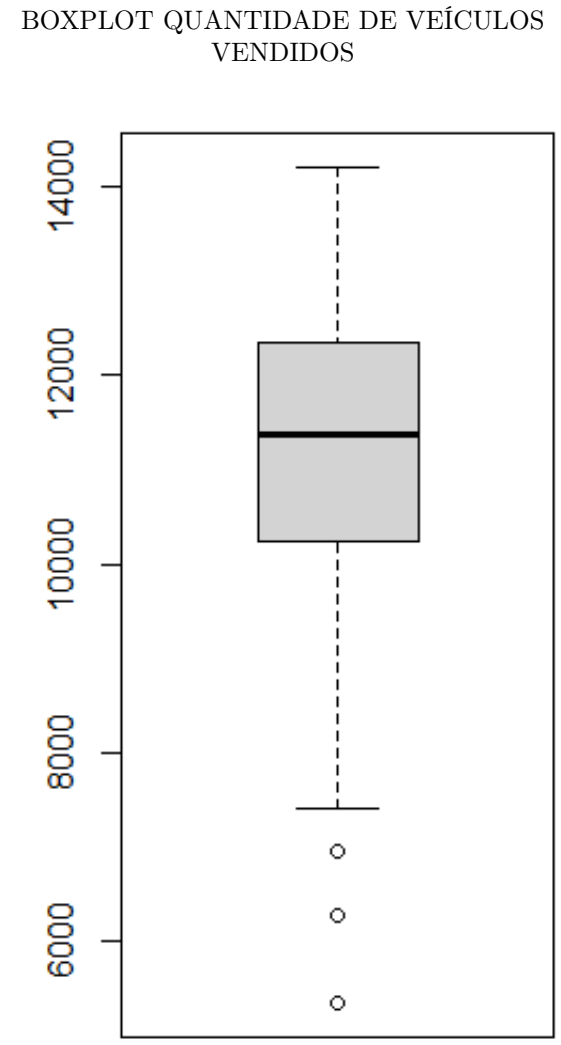
O boxplot (gráfico de caixa) é um gráfico utilizado para avaliar a distribuição empírica do dados. Tal tipo de gráfico é formado pelo primeiro e terceiro quartil e pela mediana. As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior ao limite inferior e do quartil superior até o maior valor não superior ao limite superior. Para este caso, os pontos fora destes limites são considerados valores discrepantes (outliers).

Porém ao que se refere aos quartis não será abordado nesse trabalho pois não são medidas de tendência central.

Sendo assim, com o auxílio da ferramenta R, conseguimos os seguintes BoxPlot's para cada variável analisada anteriormente (Quantidade, Quantidade & Marca e Importação) gerando boa visualização em relação às medidas de tendência central, como explicado anteriormente.

Figura 3.2: BOXPLOT VARIÁVEIS - Quantidade e Importados

((a))



((b))

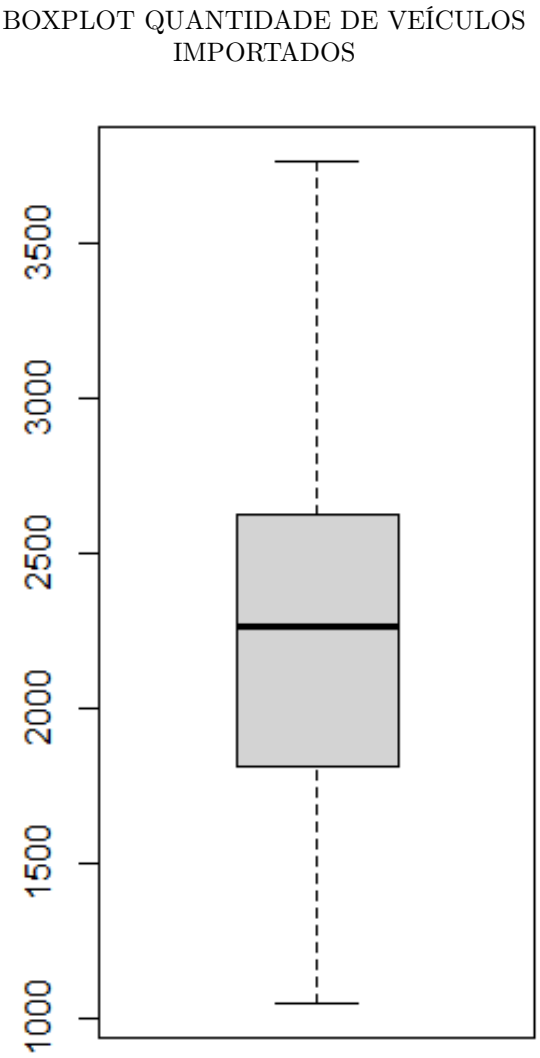
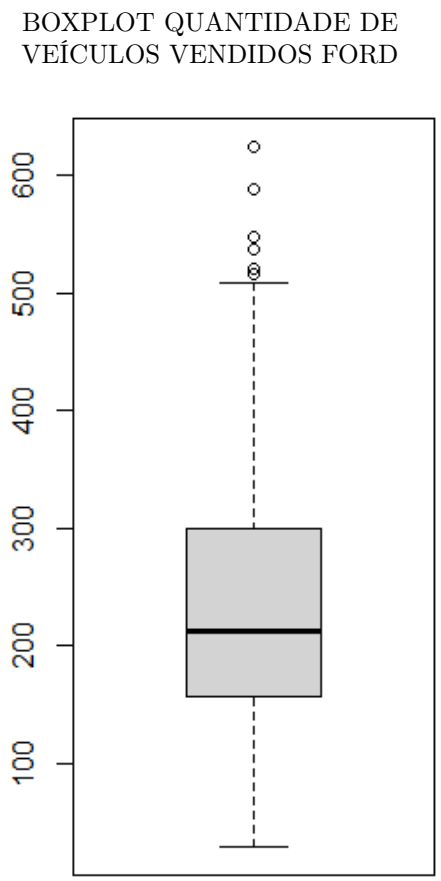
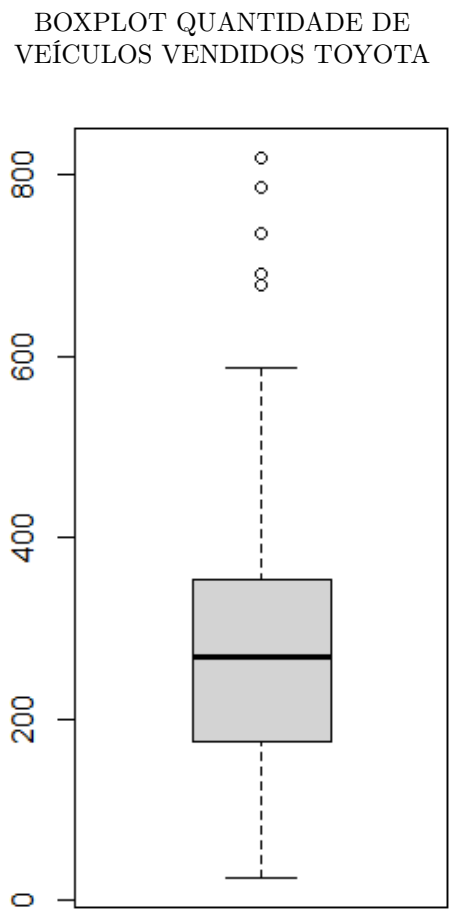


Figura 3.3: BOXPLOT VARIÁVEIS - Ford e Toyota

((a))



((b))



3.2 Medidas de dispersão

As medidas de dispersão são medidas que representam a variação dos dados numa base de dados, algumas dessas medidas são: Variância e Desvio Padrão.

Variância: A medida de dispersão denominada Variância tem a função de demonstrar o quão distante cada valor dessa base de dados está distante do valor médio. Por exemplo, quanto menor a variância, mais perto esses dados estão perto da medida central (média).

Desvio Padrão: O desvio padrão é capaz de identificar o "erro" em um conjunto de dados, caso quiséssemos substituir um dos valores coletados pela média aritmética.

3.2.1 Medidas de dispersão - Quantidade

A seguir será demonstrado o resultado das medidas de dispersão da variável ***Quantidade*** que se refere à quantidade de veículos vendidos no período do ano de 2007 até o fim de 2016.

- **Variância:** A variância, em relação à media, de quantidade de vendas de veículos na Noruega no período de ano 2007 até o ano de 2016 é de 3069667.
- **Desvio Padrão:** O desvio padrão, em relação à media, de quantidade de vendas de veículos na Noruega no período de ano 2007 até 2016 é de 1752 automóveis.

Medidas alcançadas facilmente por meio do seguinte código na linguagem R usando a ferramenta RStudio:

```
var(norway_car_month$Quantity)
3069667

sd(norway_car_month$Quantity)
1752.046
```

3.2.2 Medidas de dispersão - Quantidade & Marca

A seguir será demonstrado o resultado das medidas de dispersão da variável ***Quantidade*** em relação à variável ***Marca*** visando a quantidade de veículos vendidos em relação às marcas de veículos *Ford* e *Toyota*.

As seguintes medidas de tendência central para a variável ***Quantidade*** e ***Ford*** são:

- **Variância:** A variância, em relação à media, de quantidade de vendas de veículos da **Ford** na Noruega no período de ano 2007 até o ano de 2016 é de 11102.
- **Desvio Padrão:** O desvio padrão, em relação à media, de quantidade de vendas de veículos da **Ford** na Noruega no período de ano 2007 até 2016 é de 105 automóveis.

As seguintes medidas de tendência central para a variável ***Quantidade*** e ***Toyota*** são:

- **Variância:** A variância, em relação à media, de quantidade de vendas de veículos da **Toyota** na Noruega no período de ano 2007 até o ano de 2016 é de 15570.
- **Desvio Padrão:** O desvio padrão, em relação à media, de quantidade de vendas de veículos da **Toyota** na Noruega no período de ano 2007 até 2016 é de 124 automóveis.

Medidas alcançadas facilmente por meio do seguinte código na linguagem R usando a ferramenta RStudio:

```
#FORD
var(norway_car_model$Quantity[norway_car_model$Make == 'Ford'])
11102.27
sd(norway_car_model$Quantity[norway_car_model$Make == 'Ford'])
105.3673

#TOYOTA
var(norway_car_model$Quantity[norway_car_model$Make == 'Toyota'])
15570.52
sd(norway_car_model$Quantity[norway_car_model$Make == 'Toyota'])
124.7819
```

3.2.3 Medidas de dispersão - Importações

A seguir será demonstrado o resultado das medidas de dispersão da variável ***Importação*** que se refere à quantidade de veículos vendidos no período do ano de 2007 até o fim de 2016.

- **Variância:** A variância, em relação à media, de quantidade de importações de veículos na Noruega no período de ano 2007 até o ano de 2016 é de 328475.
- **Desvio Padrão:** O desvio padrão, em relação à media, de quantidade de importações de veículos na Noruega no período de ano 2007 até 2016 é de 573 automóveis.

Medidas alcançadas facilmente por meio do seguinte código na linguagem R usando a ferramenta RStudio:

```
var(norway_car_month$Import)
328475.6

sd(norway_car_month$Import)
573.1279
```

3.3 Formato da distribuição dos dados

O Formato da distribuição dos dados tem como missão descrever como os dados são distribuídos, explicando então se a nossa distribuição é simétrica ou assimétrica. Existindo três tipos elementares de distribuição de dados para uma variável, sendo esses três tipos nomeados como: *Assimétrica Negativa*, *Simétrica* e *Assimétrica Positiva*. Ao analisar as medidas de tendência central, poderemos dizer em qual tipo de formato citado a variável analisada irá se identificar.

Teríamos então as seguintes analises para checar se a distribuição de uma variável é simétrica ou não:

- Assimétrica Negativa: Média $<$ Mediana $<$ Moda
- Simétrica: Média = Mediana = Moda
- Assimétrica Positiva: Média $>$ Mediana $>$ Moda

Também será considerado o Índice de Assimetria de Pearson, sendo:

- Entre -0.15 e 0.15 = Praticamente Simétrica
- Entre 0.15 e 1 = Assimetria negativa
- Entre Menor que -1 ou Maior que 1 = Assimetria forte

Também é possível indicar o nível de achatamento da nossa curva, ou então, o grau de Curtose da curva de distribuição dos dados dada uma variável. Existindo três tipos elementares: *Leptocúrtica*, *Mesocúrtica* e *Platicúrtica*. Aplicando ferramentas obtidas no software R, temos as possíveis análises:

- Aproximadamente 3 = Mesocúrtica
- Menor que 3 = Platicúrtica
- Maior que 3 = Leptocúrtica

Será analisado nas variáveis anteriores (Quantidade, Quantidade & Marca, Importações) com auxílio da ferramenta R e RStudio usando os seguintes comandos:

```
install.packages("moments")
library(moments)
skewness(norway_car_month$) #Para o calculo de Assimetria.
kurtosis(norway_car_month$) #Para o calculo de Curtose.
```

Junto de funções encontradas no site universitário UFPR para visualização de dados, onde o código, e demais outro, para visualização de dados se encontra clicando aqui:

UFPR - Visualização de dados

3.3.1 Formato da distribuição dos dados - Quantidade

Como explicado anteriormente, o formato da curva de distribuição de dados é resultado da análise das medidas de tendência central para descobrir se nossa distribuição é simétrica ou assimétrica. Sendo assim temos que:

- Média: 11134
- Mediana: 11385

Sendo assim, temos que: Moda > Mediana > Média. E usando a ferramenta R, é facilmente alcançada o formato da distribuição:

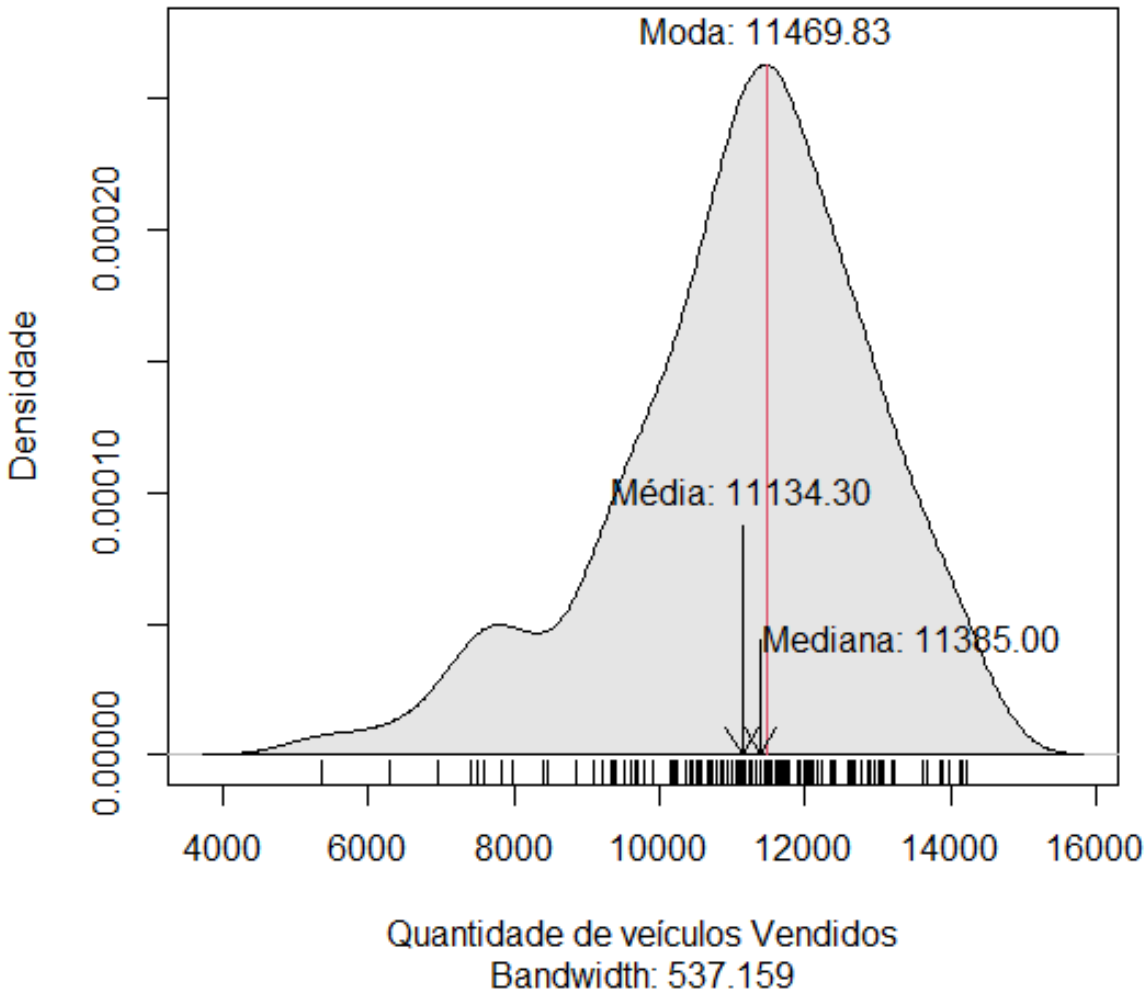


Figura 3.4: Quantidade de veículos vendidos

Portanto temos nossa distribuição é uma curva de natureza *Assimétrica Negativa* ou *Assimetria à Esquerda* de -0.7349411 e Curtose de natureza *Leptocúrtica* de 3.56929 para a variável *Quantidade* que representa a quantidade de veículos vendidos no período de ano 2007 até 2016.

3.3.2 Formato da distribuição dos dados - Quantidade & Marca

Como explicado anteriormente, o formato da curva de distribuição de dados é resultado da análise das medidas de tendência central para descobrir se nossa distribuição é simétrica ou assimétrica. Sendo assim temos que:

As seguintes medidas de tendência central para a variável ***Quantidade*** e ***Ford*** são:

- Média: 234
- Mediana: 212

Sendo assim, temos que: Média > Mediana > Moda. E usando a ferramenta R, é facilmente alcançada o formato da distribuição:

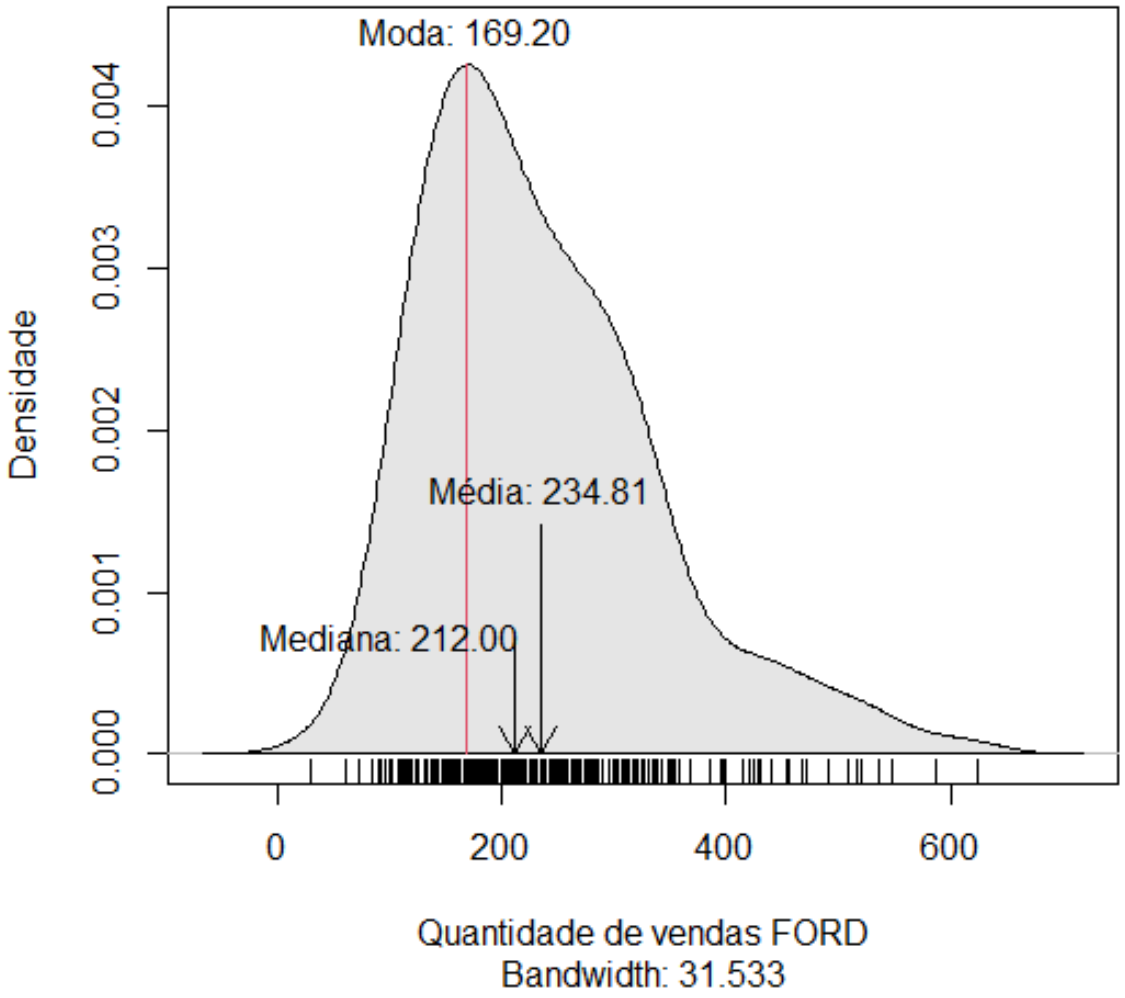


Figura 3.5: Quantidade de vendas FORD

Portanto temos que a distribuição é da natureza *Assimétrica Positiva* ou *Assimetria à Direita* de 1.019903 e Curtose de natureza *Leptocúrtica* de 4.035191 para a variável *Quantidade* em relação à marca *Ford*, que representa a quantidade de veículos vendidos no período de ano 2007 até 2016.

As seguintes medidas de tendência central para a variável ***Quantidade*** e ***Toyota*** são:

- Média: 273
- Mediana: 268

Sendo assim, temos que: Média > Mediana > Moda. E usando a ferramenta R, é facilmente alcançada o formato da distribuição:

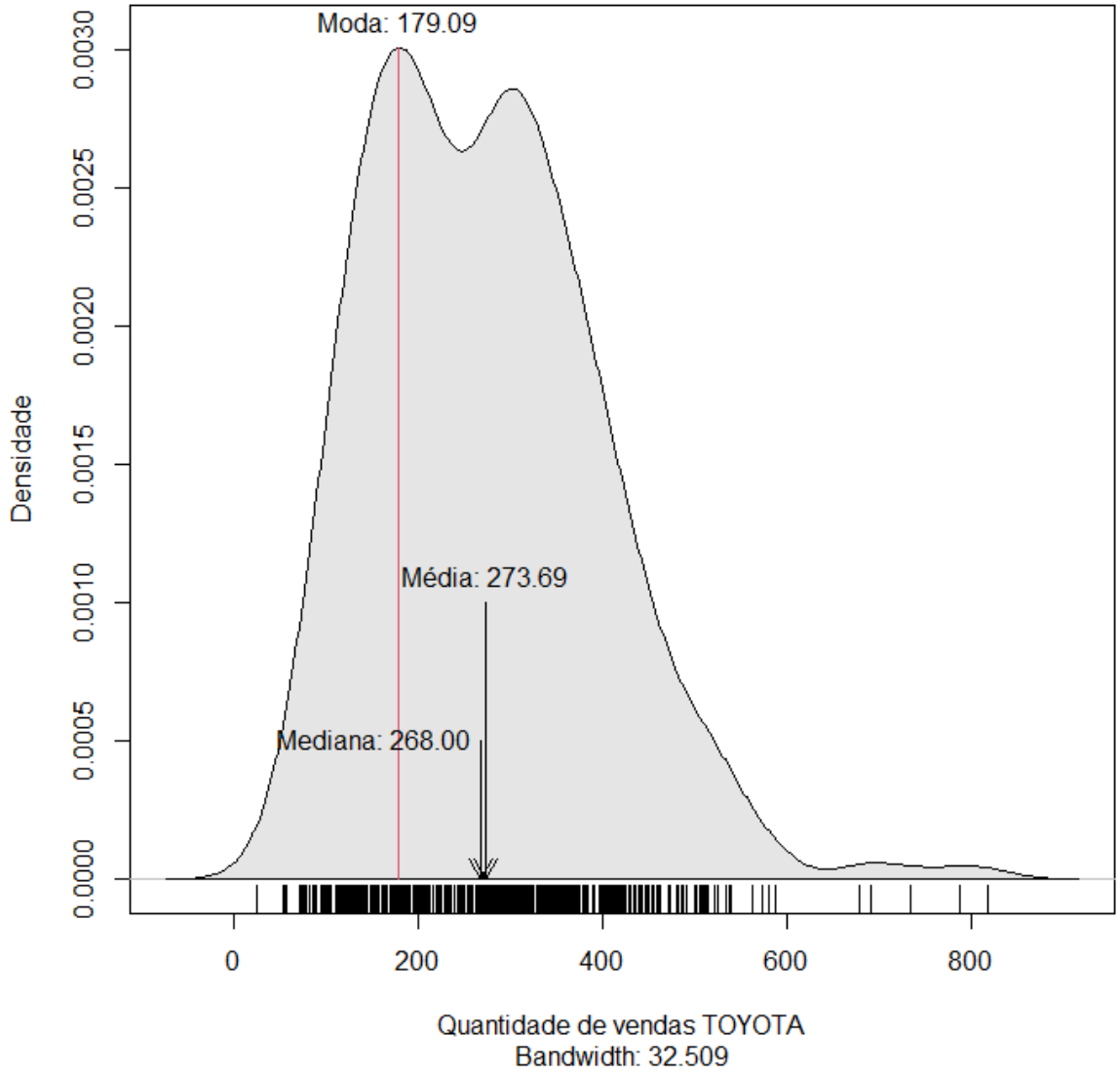


Figura 3.6: Quantidade de vendas TOYOTA

Portanto temos que a distribuição é da natureza *Assimétrica Positiva* ou *Assimetria à Direita* de 0.7423227 e Curtose de natureza *Leptocúrtica* de 3.96476 para a variável *Quantidade* em relação à marca *Toyota*, que representa a quantidade de veículos vendidos no período de ano 2007 até 2016.

3.3.3 Formato da distribuição dos dados - Importações

Como explicado anteriormente, o formato da curva de distribuição de dados é resultado da análise das medidas de tendência central para descobrir se nossa distribuição é simétrica ou assimétrica. Sendo assim temos que:

- Média: 2204
- Mediana: 2263

Sendo assim, temos que: Moda > Mediana > Média. E usando a ferramenta R, é facilmente alcançada o formato da distribuição:

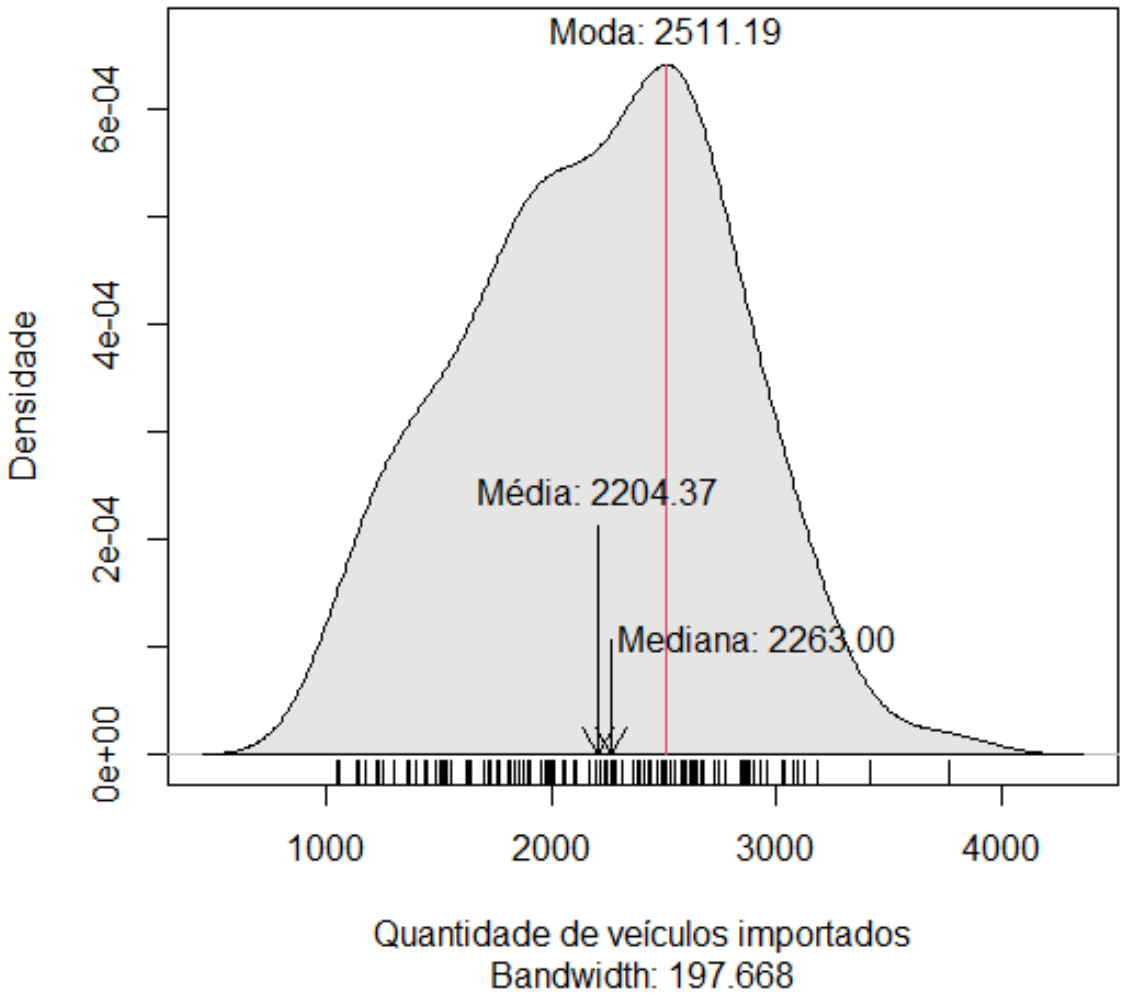


Figura 3.7: Quantidade de veículos Importados

Portanto temos nossa distribuição é uma curva de natureza aproximadamente *Simétrica* de -0.05524526 e Curtose de natureza *Platicúrtica* de 2.443223 para a variável *Importação* que representa a quantidade de veículos importados no período de ano 2007 até 2016.

3.4 Correlação entre as variáveis quantitativas

Quando possuímos variáveis quantitativas, calculamos a taxa de correlação entre elas, podendo variar de 1 até -1, sendo 1 quando as variáveis possuem correlação, -1 quando as variáveis não possuem correlação e 0 quando a correlação entre as variáveis é neutra. Com o uso de procedimentos analíticos e gráficos, podemos tornar a análise de correlação mais refinada e precisa.

3.4.1 Ideia geral - Exemplo: Quantidade & Quantidade Diesel

Uma simples análise com as seguintes variáveis: *Quantidade* e *Quantidade Diesel*, podemos verificar se a Quantidade de carros vendidos no geral possui correlação com a Quantidade de carros movidos a diesel vendidos.

Por meio de recursos gráficos na ferramenta R, é gerado o gráfico das duas variáveis, sendo a Quantidade de carros vendidos no geral no Eixo Y, e a Quantidade de carros movidos a diesel vendidos no eixo X.

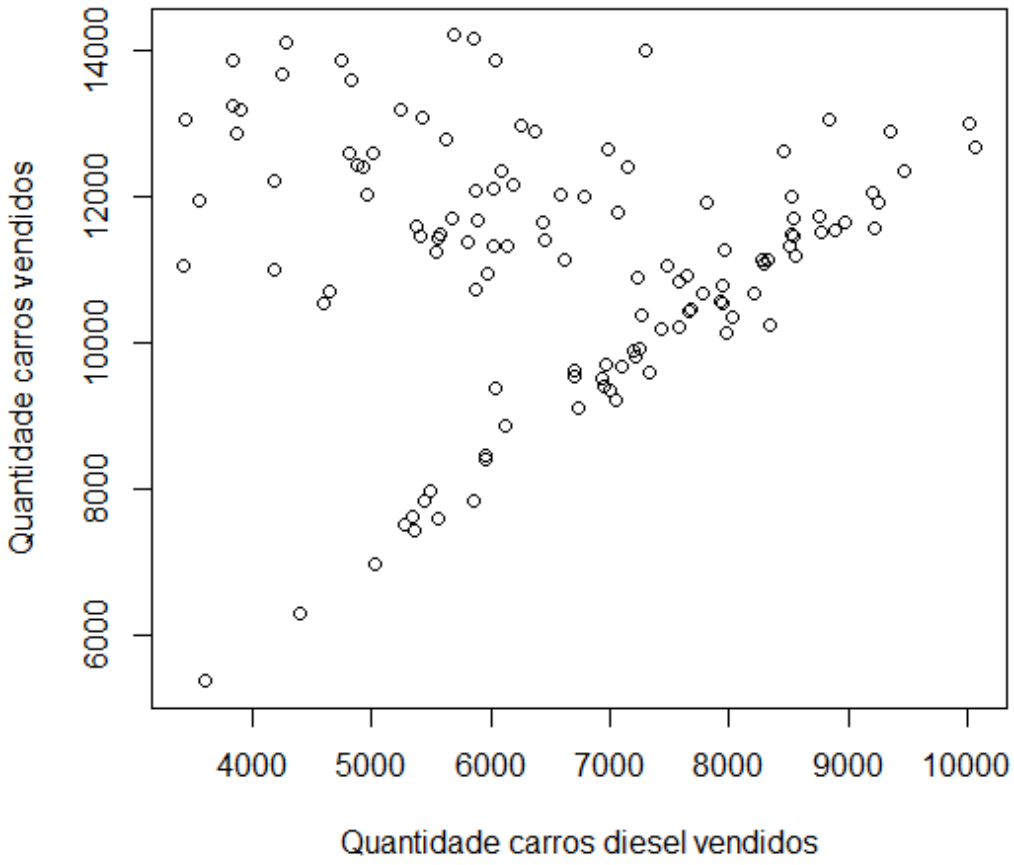


Figura 3.8: Gráfico - Quantidade vs. Quantidade Diesel

Porém ao aplicarmos a técnica *abline* para plotarmos uma linha reta, seguindo a ideia de uma regressão, temos o seguinte gráfico:

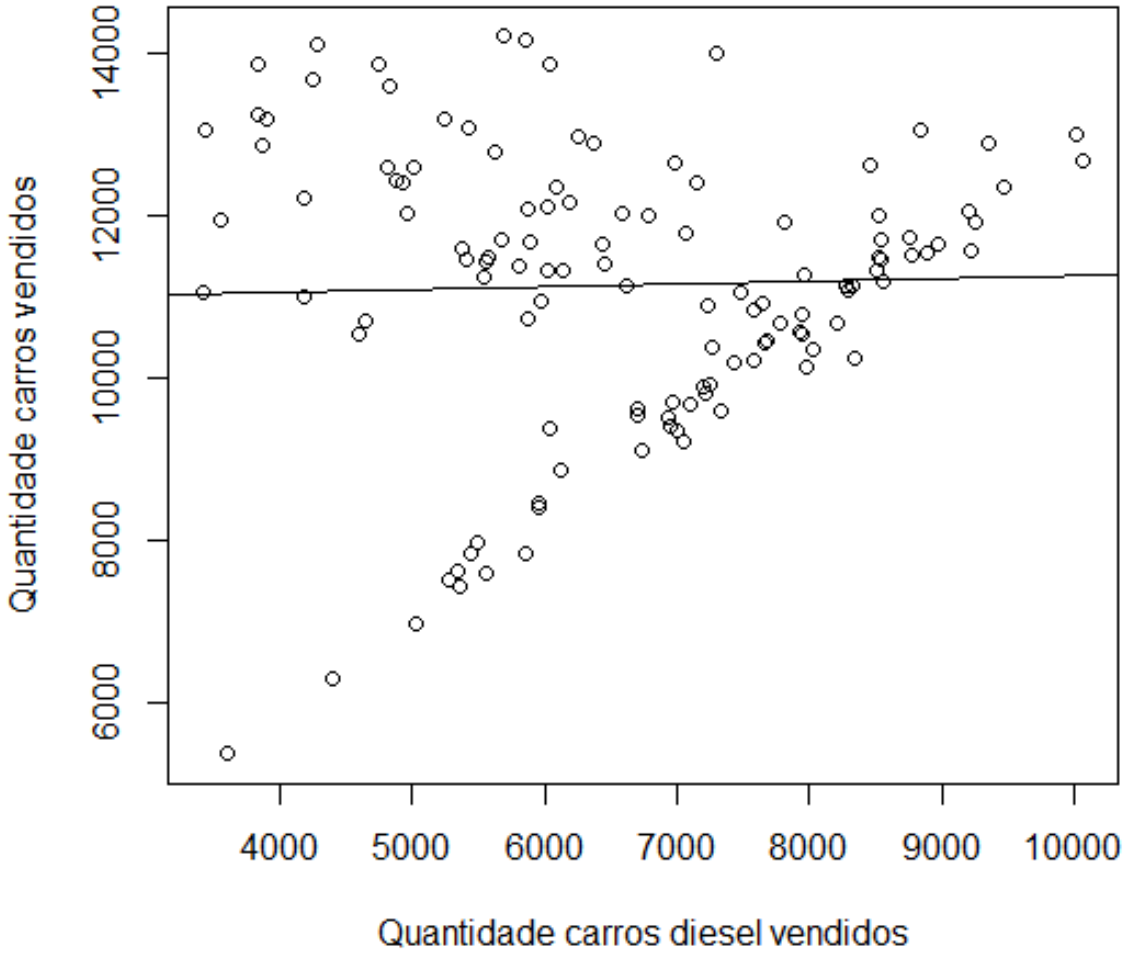


Figura 3.9: Gráfico Regressão - Quantidade vs. Quantidade Diesel

O que segure a análise de que a correlação entre as duas variáveis pode ser neutra pois a reta encontrada, de regressão linear, está quase paralela ao eixo X, o que nos geraria um coeficiente de correlação próximo de 0, porém sem certeza para afirmar.

3.4.2 Correlação das variáveis - Geral

Visando efetuar essa análise para todas as outras variáveis no banco de dados, das mais distintas maneiras, pode ser efetuado, com o auxílio da ferramenta de software R e RStudio, a seguinte análise de correlação entre as variáveis:

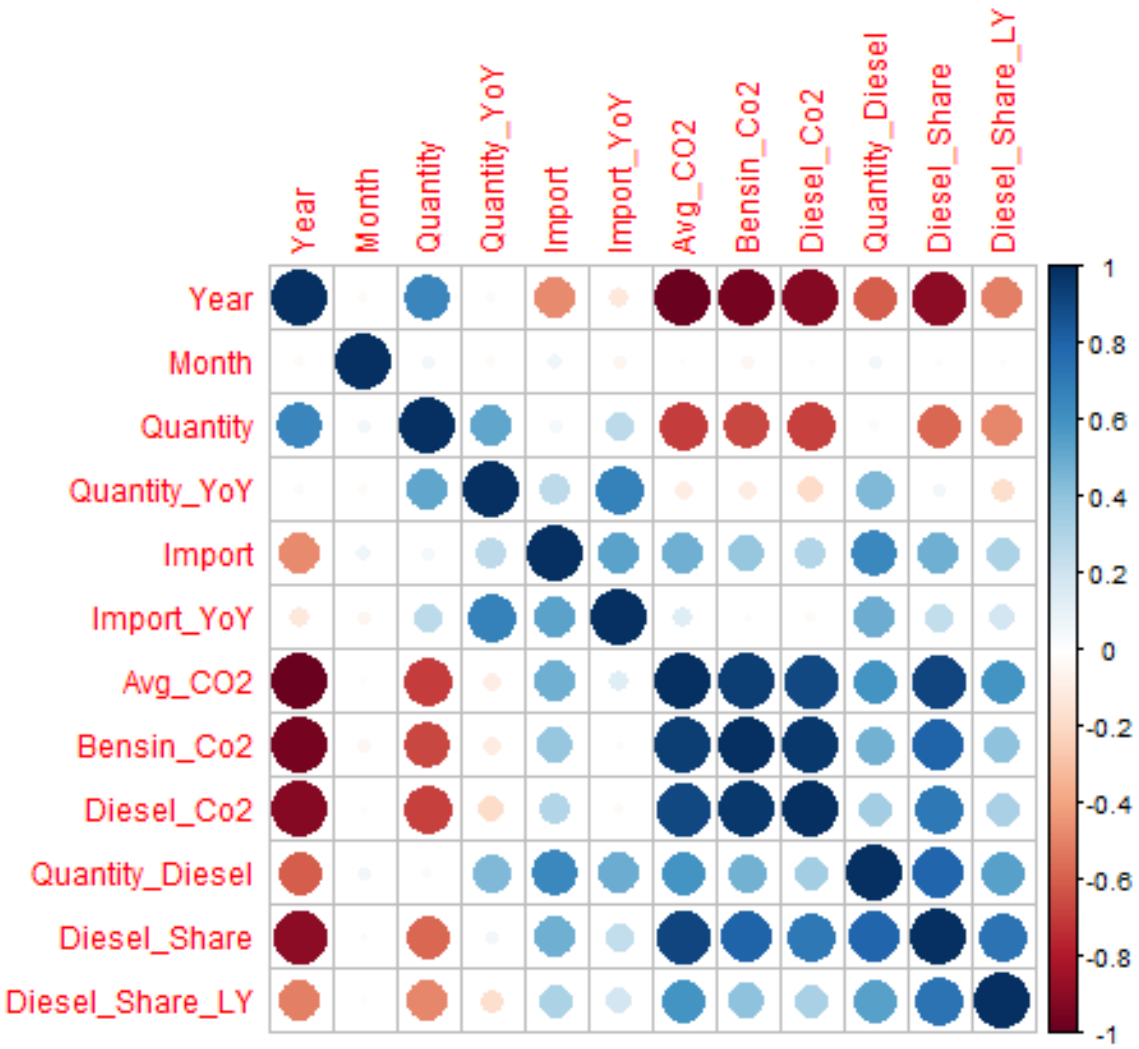


Figura 3.10: Matriz Correlação Círculo

Com a legenda devidamente posta no gráfico, temos que, quanto mais azul o encontro das variáveis na tabela, mais correlação elas possuem. E quanto mais vermelho o encontro das variáveis na tabela, menos correlação elas possuem. E quando branco significa que a correlação entre nossas variáveis é neutra.

Para uma melhor visualização das correlações podemos efetuar a mesma análise, porém ao invés de círculos, será retornado o real coeficiente de correlação:

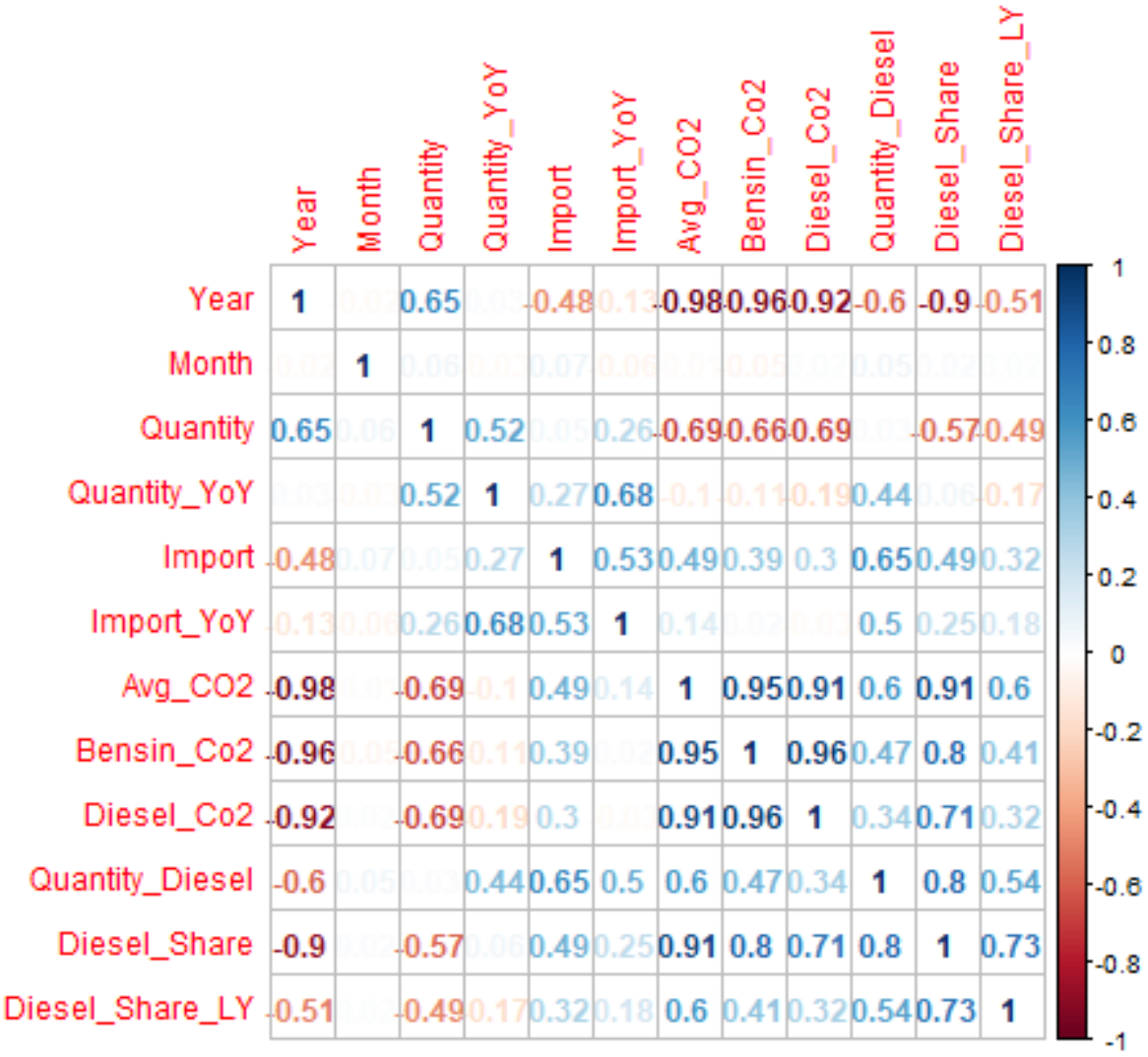


Figura 3.11: Matriz Correlação Números

Como vimos anteriormente com a plotagem do gráfico 3.9, as variáveis *Quantidade* e *Quantidade Diesel* possui correlação neutra, pois, como visualizado em nossos diagramas de correlação, o encontra das variáveis é branco, ou seja, 0 ou próximo de 0.

Podemos também agrupar tais variáveis duma maneira que fique mais fácil a visualização com o método de Cluster, resultando na seguinte imagem:

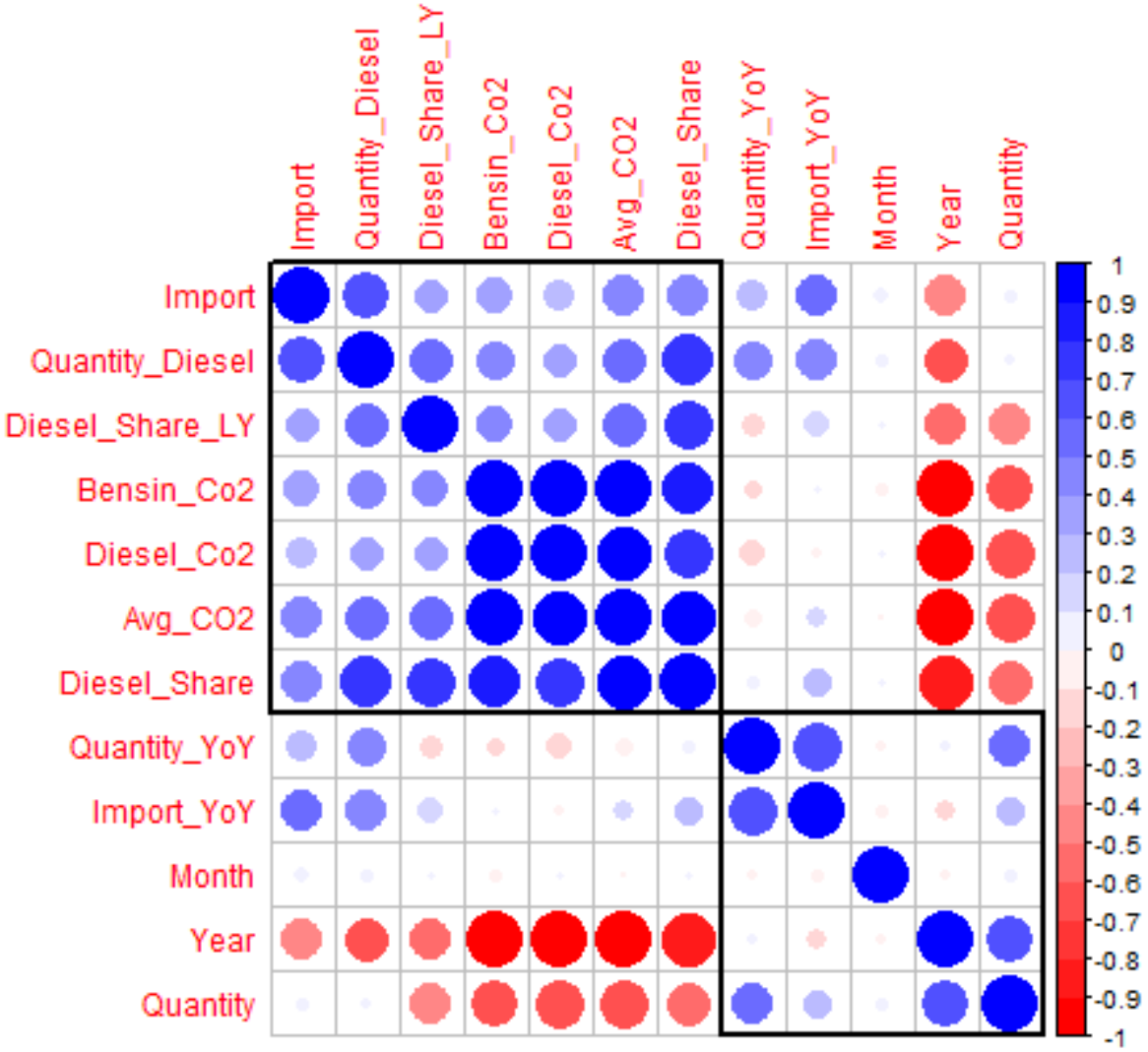


Figura 3.12: Matriz Correlação Números

Nos mostrando o resultado de uma tentativa de agrupamento de variáveis que possuíam boa, má ou neutra correlação entre si.

Dessa mesma maneira podemos fazer diversas análises de maneira simples e fácil de correlação com o software R e RStudio. Por exemplo, vemos na imagem 3.11, que as variáveis *Quantidade* e *Ano* possuem um grau de correlação agradável de 0.65, tal grau certamente significativo para uma possível análise de negócios.

Capítulo 4

Análise das medidas encontradas

4.1 Introdução às análises

Agora com as análises exploratórias de medidas de tendência central, medidas de dispersão, formato da distribuição e correlação entre as variáveis, pode-se analisá-las de maneira mais especifica para cada variável, sendo esse capítulo destinado à isso, no qual será separado tópicos para cada variável, como foi efetuado anteriormente em capítulos passados.

4.1.1 Análise - Quantidade

A quantidade de vendas, seja em qualquer negócio, é um fator muito importante e impactante, principalmente vendas que alavancam um país economicamente, como é o caso dessa base de dados do país nórdico, Noruega.

Como encontrado no capítulo anterior, temos média, mediana, variância, desvio padrão, formato de distribuição dos dados e correlação. De acordo com estudos efetuados na nossa base de dados conseguimos as seguintes medidas para a variável ***Quantidade***:

- **Média:** 11134 veículos vendidos por mês.
- **Mediana:** 11385 veículos vendidos.
- **Variância:** É de 3069667 em relação aos dados de vendas de veículos.
- **Desvio Padrão:** 1752 veículos vendidos.
- **Índice de Assimetria:** Assimetria à Esquerda de -0.7349411.
- **Curtose:** Grau de Curtose de 3.56929.

De acordo com essas medidas pode-se ver que a Média e a Mediana estão ligeiramente próximas, no que poderia nos dar a ideia de que os dados poderiam estar distribuídos de maneira simétrica, porém, ao analisar o Índice de assimetria, é evidenciado que realmente, na distribuição dos dados dessa variável possuí-se uma curva à esquerda assimétrica. Na distribuição dos dados também podemos ver um grande Grau de Curtose, que nos conclui uma distribuição assimetria à esquerda de uma curva Leptocúrtica.

Com o cálculo efetuado sobre as medidas de dispersão, é evidenciado grande variação nos dados junto com alta taxa de desvio padrão para a quantidade de vendas de veículos. Porém, um fato a ser levado em consideração (E será melhor evidenciado posteriormente com auxílio de Séries Temporais) é de que, em 2008/2009 houve uma forte crise econômica global, atingindo fortemente a venda de veículos na Noruega nessa período, tornando os dados, talvez, mais variados por esse motivo. Vemos também que tal variável possui correlação com as demais, sendo tais correlações positivas ou negativas, dificilmente são neutras, ou seja, é uma variável de suma importância. Por exemplo, como citado anteriormente que talvez o ano poderia influenciar na quantidade de vendas de veículos (Por causa da crise de 2008/2009, no exemplo citado), analisando 3.11 facilmente é visto que realmente possuem correlação pois o grau de correlação é de 0.65, o que nos faz interpretar que sim, existe correlação entre o ano e a quantidade de vendas de veículos.

Portanto, é de grande surpresa, após conseguir essas medidas, que, um país com a população de aproximadamente 5 milhões de pessoas, tenha a média de venda de veículos de **onze mil**, ou levando em consideração o desvio padrão para cima: **até treze mil** veículos, mesmo após enfrentar uma grave crise econômica global, o que nos mostra que, possivelmente, a população é economicamente ativa em relação aos seus bens materiais.

4.1.2 Análise - Quantidade & Marca

Como a análise geral de vendas de veículos está feita, podemos efetuar agora a análise em cima de algumas marcas automobilísticas famosas, como Ford e Toyota.

As seguinte medidas foram conseguidas através de analises na variável ***Quantidade*** & ***Ford***:

- **Média:** 234 veículos vendidos por mês.
- **Mediana:** 212 veículos vendidos.
- **Variância:** É de 11102 em relação aos dados de vendas de veículos.
- **Desvio Padrão:** 105 veículos vendidos.
- **Índice de Assimetria:** Assimetria à Direita de 1.019903.
- **Curtose:** Grau de Curtose de 4.035191.

As seguinte medidas foram conseguidas através de analises na variável ***Quantidade*** & ***Toyota***:

- **Média:** 273 veículos vendidos por mês.
- **Mediana:** 268 veículos vendidos.
- **Variância:** É de 15570 em relação aos dados de vendas de veículos.
- **Desvio Padrão:** 124 veículos vendidos.
- **Índice de Assimetria:** Assimetria à Direita de 0.7423227.
- **Curtose:** Grau de Curtose de 3.96476.

Após a conclusão de todas as medidas para os dois tipos de marca, é facilmente visto que, ligeiramente, a marca *Toyota* obteve melhores resultados em quesito de venda de automóveis em relação à outra marca analisada, *Ford*. Sendo assim, talvez a marca *Ford* deveria rever sua estratégia de vendas nesse país para poder alavancá-los e vencer um de seus principais concorrentes com números expressivamente melhores do que foram encontrados nas análises feitas desse trabalho durante o ano de 2007 até 20016. Certamente tal investimento em estratégia de venda faria muito sentido, pois como visto anteriormente, a Noruega é um país em que a compra de carros pela população ocorre com grande frequência e com grande certeza é um mercado do qual vale a pena o investimento para dominar o mercado de vendas automobilísticas.

4.1.3 Análise - Importação

A Noruega, de acordo com o site CEIC, disponível em:

CEIC - Importações Geral Noruega

Pode-se verificar que a Noruega é um país do qual a parte econômica de importação (No geral) é muito ativa e comumente em alta. Visando isso, apresentamos as medidas encontradas para a variável ***Importação*** nos anos de 2007 até o fim de 2016:

- **Média:** 2204 veículos importados por mês.
- **Mediana:** 2263 veículos importados.
- **Variância:** É de 328475 em relação aos dados de vendas de veículos.
- **Desvio Padrão:** 573 veículos vendidos.
- **Índice de Assimetria:** Simétrica de -0.05524526.
- **Curtose:** Grau de Curtose de 2.443223.

Como é evidenciado pelo portal CEIC, pode-se ver que a taxa geral de importação é positiva durante esse período de quase dez anos que foram analisados nesse trabalho, sendo então, a Noruega, um país que realmente trabalha com importações em seu setor econômico, e possivelmente também, no setor econômico automobilístico. Além disso, como visto na imagem 3.7, vemos que o grau de assimetria é de -0.05, que, de acordo com o Índice de assimetria de Pearson, quando entre -0.15 e 0.15, podemos considerar a distribuição simétrica. Então, é indicado que a concentração da quantidade de importação não foi tão alta quanto poderia ser, porém bem concentrada, o que é evidenciado na média e mediana, que são bem próximas.

Também, podemos ver na figura 3.2(b) que possuímos números baixos, em relação à média, de importações em alguns meses, sendo o mínimo próximo de mil importações por mês, talvez também resultado da crise de 2008/2009.

Capítulo 5

Regressões e Previsões

5.1 Regressão Linear

O conceito de regressão tenta ajudar a prever comportamentos com base na associação entre duas variáveis que geralmente possuem uma boa correlação. Existem diversos tipos de regressões, porém nesse trabalho será consumido apenas a regressão de tipo Linear. A Regressão Linear, como seu próprio nome diz, trabalha com funções lineares e tentará confeccionar uma reta que tente passar o mais próximo de cada ponto existente no plano cartesiano após a plotagem de duas variáveis em cada eixo do plano.

Por exemplo, como vimos na imagem 3.9 para tentar verificar inicialmente se, com o comportamento da reta, existia correlação entre duas variáveis. Com o auxílio do portal da Universidade Federal do Paraná para a criação de gráficos mais refinados agradavelmente visíveis, foi possível análises de regressões interessante.

5.2 Regressão Linear - CO2 vendidos vs. Ano

Como foi citado anteriormente, a Noruega nos dias de confecção desse trabalho (2020), se tornou uma grande potência no quesito carros elétricos (Porém como a base de dados vai até 2016 não pode ser feito tal análise pois possuíamos muitos dados faltantes, como visto no Capítulo 2). Uma maneira de verificarmos a queda de carros tradicionais é analisarmos o inverso do que seria um carro movido à eletricidade, sendo assim, analisaremos carros vendidos que possuem emissão de CO2.

Figura 5.1: Regressão Linear - CO2 vendidos vs. Ano

Com a regressão podemos ver que sim, ao passar dos anos, a partir de 2013 houve queda drástica e contínua em relação aos carros vendidos que emitem CO2. Em contra partida analisaremos agora a quantidade de carros vendidos por ano, tentando verificar se a quantidade de carros vendidos por ano aumentou, se sim, nos mostra que a população está abandonando cada vez mais o modelo de carro tradicional emissor de CO2.

5.3 Regressão Linear - Quantidade vendidos vs. Ano

Aplicaremos agora a tentativa citada anteriormente, tentaremos ver se a quantidade de carros vendidos cresceu ao passar dos anos. Com o auxílio do software R e RStudio chegamos ao seguinte gráfico de Regressão Linear para as variáveis *Quantidade* vs. *Ano*:

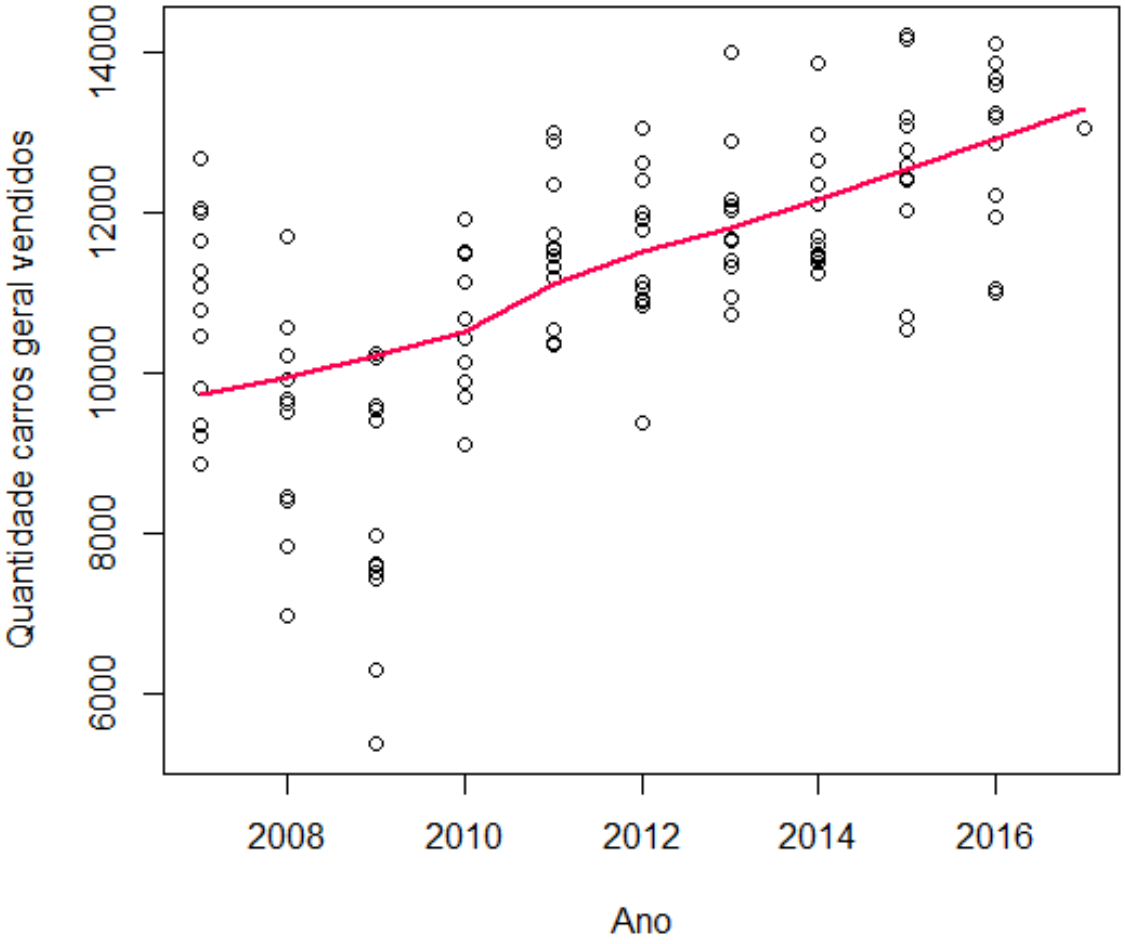


Figura 5.2: Regressão Linear - Quantidade geral veículos vendidos vs. Ano

Como era esperado após análises anteriores, a Quantidade de carros no geral aumentou ao passar dos anos. E, outro fator interessante, é que pode ser visto que no período de crise global, 2008/2009, houve grande queda porém em um curto período essa parte da economia voltou a crescer na Noruega.

Portanto, pode-se dizer que a quantidade de carros emitentes de CO2 diminuiu mas em contra partida a Quantidade de carros no geral aumentou com o passar dos anos, nos levando à ideia de que realmente a Noruega é um país potência em relação à vendas de carros híbridos e/ou elétricos. Talvez isso nos mostre o porquê a marca *Ford* não possuiu tanto sucesso em relação às vendas de automóveis do que a marca concorrente, *Toyota*. *Ford* não possui carros elétricos para serem vendidos e acaba por a marca não se inserir tão bem quanto a concorrente *Toyota* faz, pois a marca *Toyota* possui carros elétricos à venda no mercado.

5.3.1 Série Temporais

No ramo dos negócios, comumente é levado em consideração as vendas conquistadas em cada mês para cada ano, uma ferramenta que nos auxilia muito quanto à isto é a ferramenta de Séries Temporais. Com o uso de Séries temporais, poderemos ver o geral obtido em cada ano, ou até melhor, as vendas efetuadas em cada mês de cada ano, abrindo caminho assim para uma possível previsão usando modelos de previsões.

5.3.2 Série Temporal - Quantidade vs. Ano

Plotando uma Séria Temporal geral, com auxílio da ferramenta R e RStudio, conseguimos chegar à seguinte imagem:

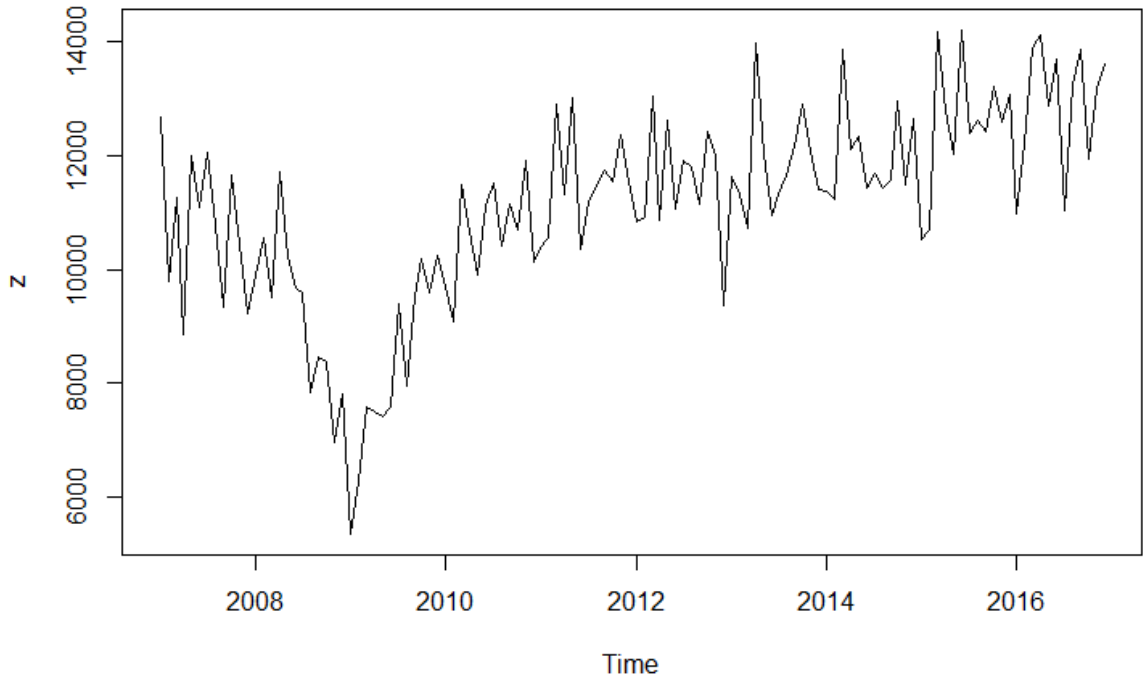


Figura 5.3: Série Temporal - Quantidade vendidos geral vs. Ano

E agora, com toda certeza vemos, no período de crise global, 2008/2009, houve uma grande queda nas vendas de veículos na Noruega por conta de tal acontecimento. Podemos ver, inicialmente, sazonalidade nessa Série Temporal, o que, posteriormente, certamente poderá influenciar o modelo de previsão.

5.3.3 Série Temporal - Quantidade vs. Mês e Ano

Para analisarmos em cada mês de cada ano, podemos separar da seguinte maneira:

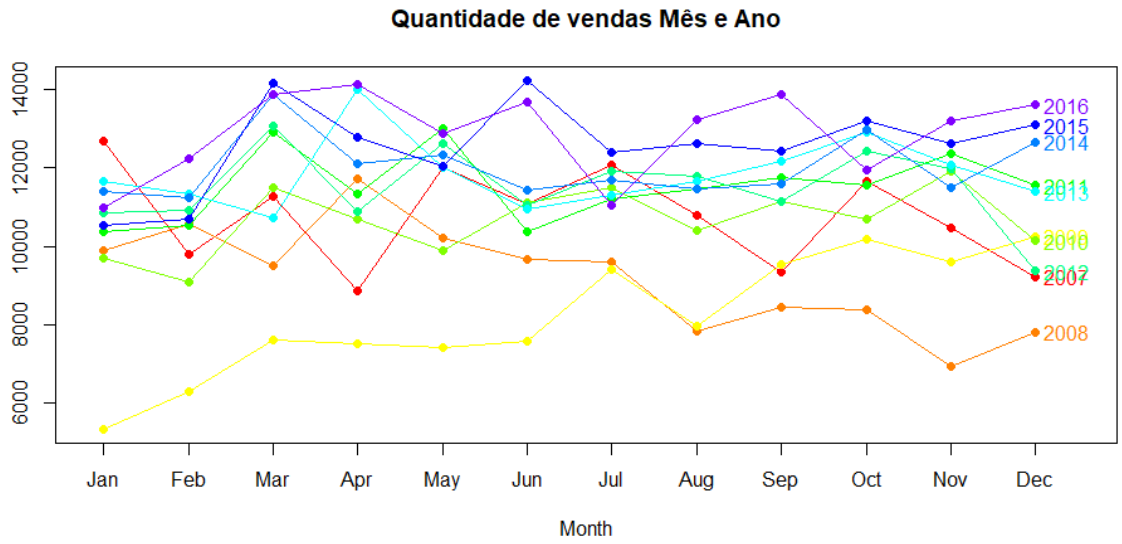


Figura 5.4: Série Temporal - Quantidade vendidos vs. Mês e Ano

Dessa maneira podemos fazer uma análise cirúrgica para cada ano. Em 2007 por exemplo, houve grande variação nos dados de cada mês, possivelmente uma má fiscalização sobre a variável Quantidade de veículos vendidos, uma vez que a coleta desses dados era iniciada nesse ano. Outro breve exemplo, em 2008, no inicio da crise houve grande queda até o fim do mesmo ano, iniciando 2009 com grande queda ainda e se recuperando mesmo apenas no segundo semestre de 2009. Nos revelando a grande recuperação econômica de uma das potências mundiais, como é a Noruega.

Modelos facilmente concluídos com ajuda do software R e RStudio com o seguinte código:

```
install.packages("forecast")
library(forecast)

#Serie Temporal (QUANTIDADE DE CARROS VENDIDOS NOS ANOS)
vetor <- as.numeric(norway_car_month$Quantity)
z <- ts(vetor, frequency = 12, start = c(2007,1),
        end = c(2016,12))

#Quantidade geral de carros vendidos
ts.plot(z)

#Quantidade mes a mes em cada ano de carros vendidos
seasonplot(z, col = rainbow(12), border = c("royalblue"),
           year.labels = TRUE, Type = "o", pch= 16,
           main = "Quantidade de vendas Mes e Ano")
```

5.3.4 Previsão

Com o modelo temporal já finalizado, podemos efetuar a previsão para os próximos quarenta e quatro meses usando o modelo de previsão no software R chamado Holt-Winters. Porém iremos levar em consideração as tendências e possíveis ajustes sazonais.

Sendo assim, após efetuar a finalização do modelo Holt-Winters, teríamos:

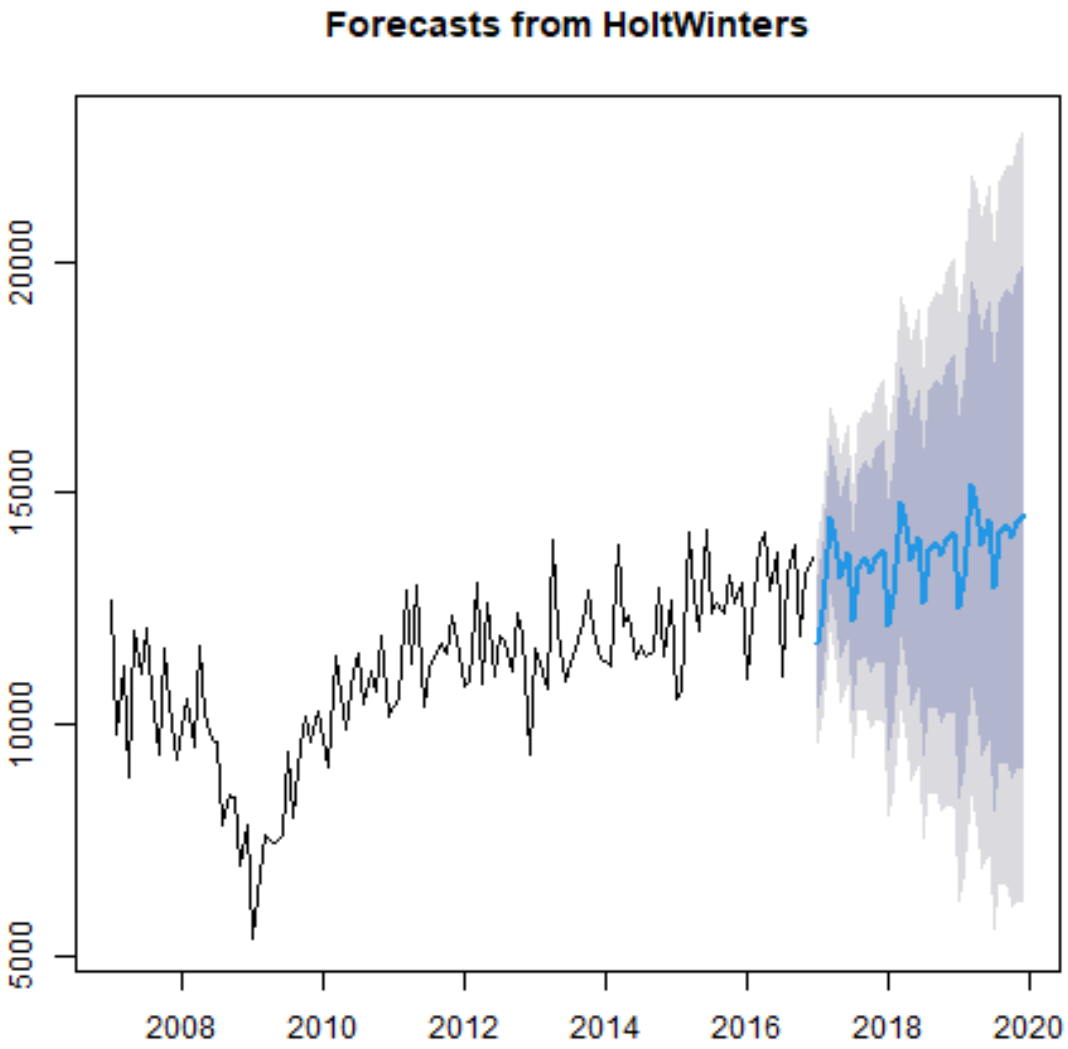


Figura 5.5: Série Temporal - Quantidade vendidos vs. Mês e Ano

O modelo gerado tentou prever-nos até o ano de 2020, nos indicando um ligeiro aumento nas Quantidades de vendas de veículos. Sendo assim, a silhueta azul escuro se refere à possíveis desvios padrões em relação à reta criada com 80% de certeza, seja para cima ou para baixo. Já a silhueta cinza se refere à possíveis desvios padrões em relação à reta criada com 95% de certeza, seja para cima ou para baixo.

O modelo de previsão Holt-Winters foi obtido a partir do seguinte código em plataforma R:

```
#Previsao com Holt-Winters com Tendencia e Ajuste sazonal
ajuste_HW_CT_CS <- HoltWinters(z)
ajuste_HW_CT_CS
plot(ajuste_HW_CT_CS)

previsao <- forecast(ajuste_HW_CT_CS, h=36)
plot(previsao, main = "Previsao Venda de Carros Noruega 2017 - 2019",
        xlab = "Anos",
        ylab = "Quantidade de Carros Vendidos")
```

5.3.5 Provando a previsão encontrada

Como explicado na introdução, nossa base de dados possui o dado de janeiro de 2017, porém foi descartado nas análises, o número de Quantidade de vendas de Janeiro de 2017 foi de 13055 veículos vendidos.

Chamando a variável "Previsão"teremos retornado todos as previsões feitas para cada mês, como por exemplo (Sendo Lo = Low % e Hi = High %):

```
> previsao
         Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
Jan 2017       11768.32 10336.662 13199.98  9578.788 13957.85
Feb 2017       12430.94 10920.643 13941.24 10121.140 14740.74
Mar 2017       14445.68 12854.275 16037.09 12011.835 16879.53
```

Sendo assim, nosso modelo previu com 80% de certeza para cima (Hi 80) que aconteceria aproximadamente 13200 vendas de veículos, número bem próximo do que realmente ocorreu na realidade, do qual aconteceram 13055 veículos vendidos.

Capítulo 6

Conclusão

Dessa maneira podemos concluir, após as análises feitas, que a Noruega possui um grande mercado de vendas de automóveis, tal mercado que está se moldando para apenas veículos híbridos e ou elétricos, abandonando os veículos tradicionais.

Também foi possível verificar que a grande economia norueguesa auxiliou para que o país enfrentasse a crise econômica de 2008/2009 no setor automobilístico e em 2010 essa parte da economia do país ja voltasse a crescer novamente.

Foi verificado que, caso não haja nenhuma outra crise, global ou interna, a Noruega continuará tendo um fértil mercado de automóveis, porém automóveis que sigam a tendência que foi analisada, ou seja, carros elétricos. Marcas e Importações de veículos que não seguirem essa tendência, poderão enfrentar fracasso em relação às vendas de seus veículos.

Referências Bibliográficas

Crawley(2012) Michael J Crawley. *The R book*. John Wiley & Sons. Citado na pág.

Morettin e BUSSAB(2017) Pedro Alberto Morettin e WILTON OLIVEIRA BUSSAB. *Estatística básica*. Saraiva Educação SA. Citado na pág.

Provost e Fawcett(2013) Foster Provost e Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.". Citado na pág.