|  |  |
| --- | --- |
| **FACULDADE DOM BOSCO DE PORTO ALEGRE**  Credenciada pela Portaria Nº 3.254, de 26 de Novembro de 2002  **Curso de Sistemas de Informação**  (Reconhecido pela Portaria 1093 de 24/12/2015, publicado no DOU de 30/12/2015) | |
| **Disciplina: Introdução a Ciência de Dados** | |
| **Valor: 30%** | **Data: 01/12/2022** |
| **Professor: Adriana Paula Zamin Scherer** | |
| **Projeto Final da Disciplina – Parte 1** | |
| **Aluno: Marcos Gabriel Koslovski Santos** | |

Regressão linear simples e multipla de um dataset de veículos usados:

A base escolhida no site <https://www.kaggle.com/> contém informações sobre carros usados como nome, ano, valor de venda, quilometragem, combustível, tipo de vendedor, transmissão e usuário.

O referido dataset foi construido a partir de um site de vendas de veículos usados indiano chamado de Car Dekho, portanto, pode ajudar o usuário a encontrar respostas referentes ao mercado de veículos na Índia como relações entre quilometragem e combustível com valor de venda, entre outras.

Sabemos de antemão que a idade é uma variável determinante no valor de um automóvel e esperamos encontrar no dataset através de regressão linear este resultado.

No dataset temos as seguintes colunas:

* name: nome, modelo e fabricante (por vezes aparece versão, quantidade de assentos e outras informações relevantes);
* year: ano de fabricação do automóvel;
* selling\_price: valor de venda do automóvel;
* km\_drive: kilometragem do automóvel;
* fuel: combustível que o automóvel consome;
* seller\_type: tipo do vendedor que está vendendo o automóvel;
* transmission: tipo de transmissão do automóvel (manual/automático);
* owner: quantidade de donos do automóvel.

**Problema central:**

A partir do enunciado do trabalho proposto pela professora Adriana Scherer iniciamos a busca por um dataset adequado com o que tínhamos em mente para realização da tarefa proposta. Num primeiro momento escolhemos um dataset incompatível com o desenvolvimento do trabalho, onde com o apoio de Adriana então encontramos um dataset adequado que tem como conteúdo os dados de diversos automóveis usados que estão a venda.

A intenção deste trabalho é encontrar através de regressão linear simples a relação entre o valor de venda de um automóvel e sua idade.

Através de regressão linear múltipla esperamos encontrar a relação entre o valor de venda, quilometragem e a idade.

**Resultados encontrados:**

**1) Análises quantitativas:**

Média de ano de fabricação: 2014

Maior ano de fabricação: 2020

Menor ano de fabricação: 1992

Maior valor de venda: 8.900.000,00

Menor valor de venda: 20000,00

Média de valor anunciado: 503.928,30

Total dos carros: 4342

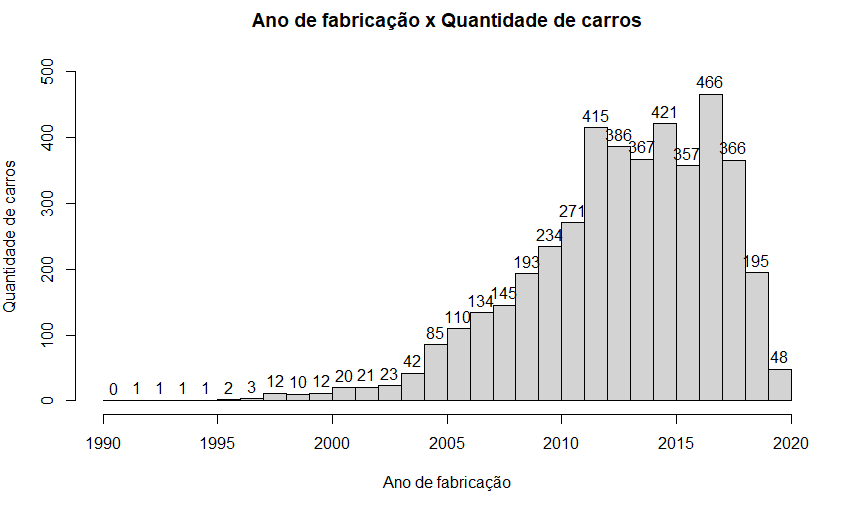
Quantidade de carros por ano:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1992** | **1993** | **1994** | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** |
| 1 | 1 | 1 | 1 | 2 | 3 | 12 | 10 | 12 | 20 |
| **2002** | **2003** | **2004** | **2005** | **2006** | **2007** | **2008** | **2009** | **2010** | **2011** |
| 21 | 23 | 42 | 85 | 110 | 134 | 145 | 193 | 234 | 271 |
| **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** | **2020** |  |
| 415 | 386 | 367 | 421 | 357 | 466 | 366 | 195 | 48 |  |

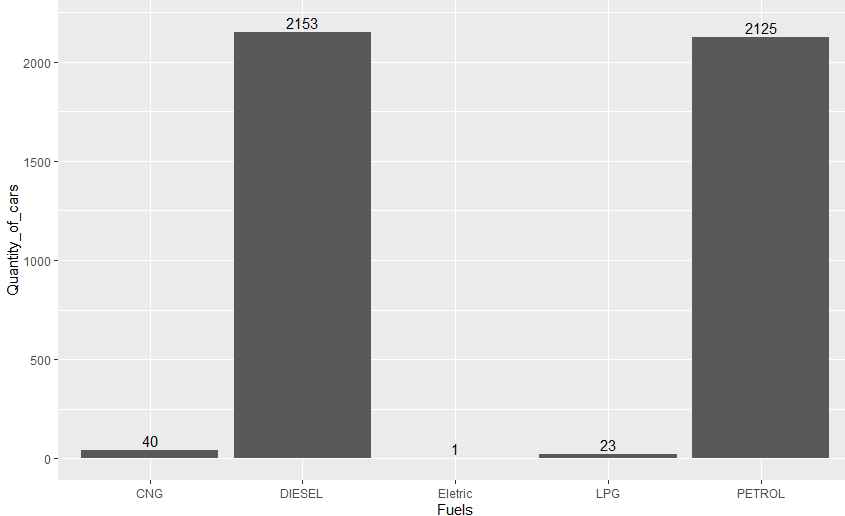
Quantidade de carros por combustível:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CNG** | **DIESEL** | **ELETRIC** | **LPG** | **PETROL** |
| 40 | 2153 | 1 | 23 | 2125 |

Histograma quantidade de carros por ano:



Barplot da quantidade de carros por combustível que consomem



Resumo:

Ao final das análises quantitativas verificamos grandes discrepâncias entre valores mínimos e máximos de praticamente todas as varíaveis. Este era um sinal de alerta para a parte do trabalho usando machine learning já que desde já não identificamos mesmo que visualmente padrões nos dados do dataset.

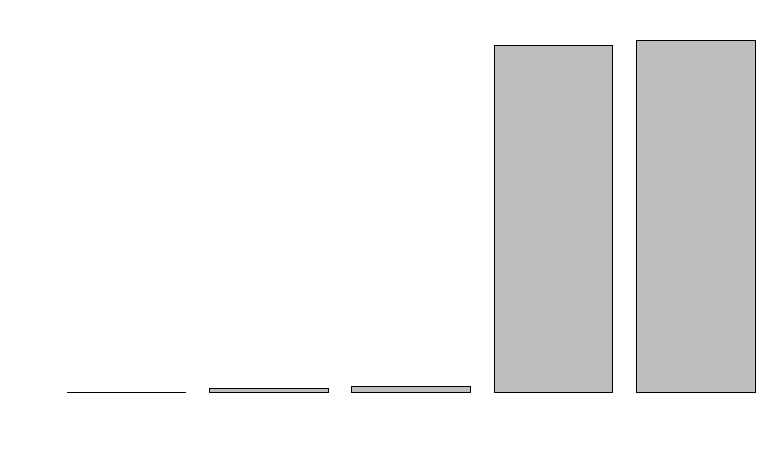
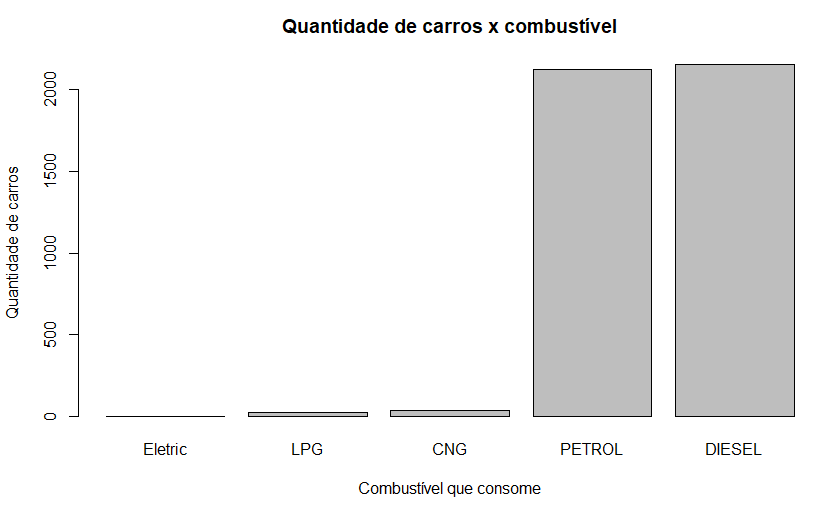
A grande variância dificulta a análise de valores médios de grandezas numéricas já que, por exemplo:

* Somente 3.431% dos veículos são mais velhos do que o ano de 2005;
* Somente 1,473% dos veículos são movidos por outros combustíveis além de diesel e gasolina;
* Total de 49.585% dos veículos são movidos a diesel;
* Total de 48.940% dos veículos são movidos à gasolina;
* Diferença de 28 anos entre o veículo mais novo e o veículo mais velho;
* Diferença de 8.880.000,00 entre o maior e o menor valor de veículo;
* Menor valor representa somente 0,2247191% do maior valor
* 1491 descrições de veículos diferentes
* O veículo Maruti Swift Dzire VDI com sua descrição unica possui 69 anuncios
* 695 descrições únicas com somente um veículo anunciado.

Encontramos dificuldades quando estavamos realizando o bar plot da quantidade de carros por combustível por combustível. Sabendo que era importante as quantiades por escrito tentamos setar a label como verdadeiro ao invés de aparecerem às quantidades sumiam todas as labels do grafico e o seguinte erro aparecia:

"labels" não é um parâmetro gráfico

O gráfico que antes não apresentava a quantidade por barra agora já não apresentava dado nenhum conforme abaixo:



Cientes do problema fizemos uma busca onde descobrimos que não é possivel mostrar as quantidades em um gráfico construido via barplot. Buscamos outra biblioteca em que fosse possível realizar o trabalho e encontramos o gnuplot2, que ainda que sob um padrão um pouco diferente nos mostrou o gráfico exatamente como esperavamos. É por isso que o histograma da quantidade de carros por ano é um pouco diferente do barplot da quantidade de carros por combustível que consomem.

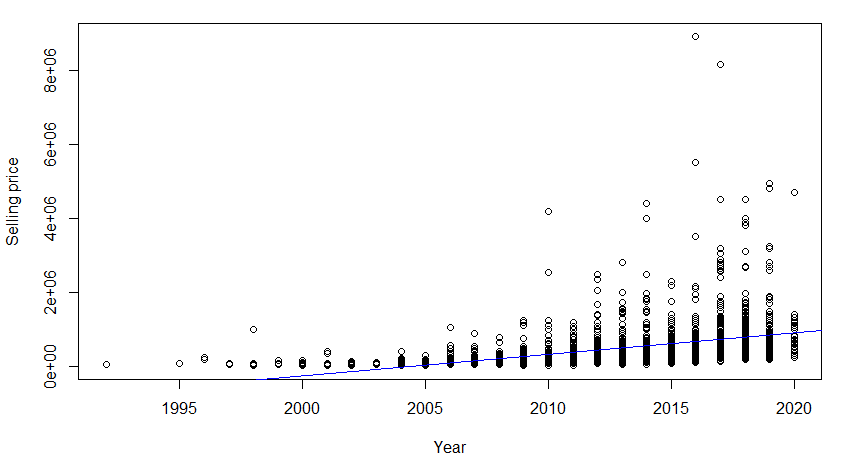
2) Machine learning :Regressão linear simples: Valor de venda x idade

Correlação: 0,41

Média diferença: 263.492,60

Acuracia: 0.1798569

Erro residual: 541.100,00 em 3065 graus de liberdade



3) Machine learning :Regressão linear multipla:

Média diferença: 263.640,808

Acuracia: 0.1802744

Erro residual: 541100 em 3064 graus de liberdade

Resumo:

Ao final das análises referentes a regressão linear simples e multipla verificamos que ambas apresentaram baixissima acurácia e correlação. A média em ambas as regressões foi altissima, maior inclusive que o valor de muitos automóveis do dataset sendo impossivel utilizar os retornos do aprendizado para quaisquer fins.

Assim como a analise quantitativa a grande diversidade de valores e a baixa correlação entre as variaveis afetaram duramente os resultados das regressões. Isso em linhas gerais significa que usando o dataset completo não necessáriamente um carro pouco rodado é mais caro que um muito rodado ou um carro mais novo é mais caro que um carro mais velho.

O resultado poderia ser diferente se avaliassemos por modelo de automóvel, por exemplo, porém um número limitado de automóveis a venda de mesma marca e modelo no dataset também afetaria os resultados.

Resumidamente o grande problema está no dataset escolhido. o numero limitado de informações referente aos automóveis e a grande variancia entre kilometros rodados, ano de fabricação e valor e modelo

Obs: Testamos também a correlação utilizando de regressão linear simples entre kilometragem e valor de venda e a correlação foi ainda pior, chegando a ser negativa.

Bônus!

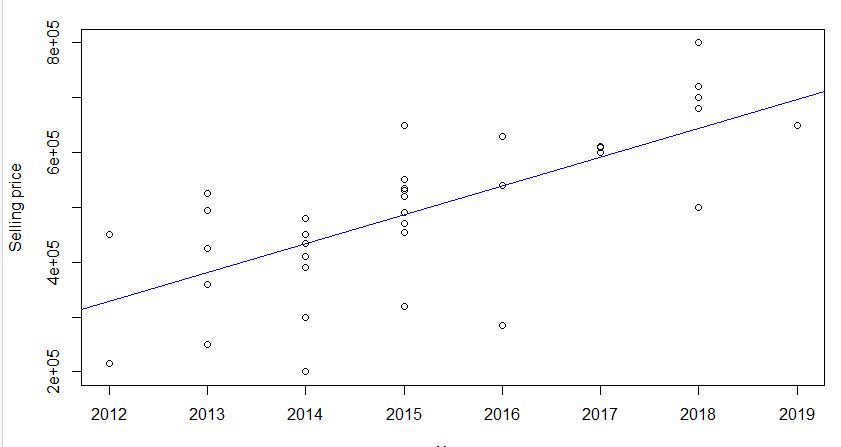
Geramos um dataset somente com o veículo que temos a maior quantidade de anúncios, o Maruti Swift Dzire VDI, e com os dados de 69 registros chegamos aos seguintes resultados:

Correlação: 0.7614072

Média diferença: 65299.1

Acuracia: 0.5888623

Erro residual: 89050 em 46 graus de liberdade



Resumo:

Como informado anteiormente numa amostra menor e se tratando do mesmo veículo tivemos resultados muito melhores do que usando todo o data set.

Acredito assim termos valores que podem sim ser usados para previsão de valores já que ficou clara a correlação entre a idade e valor de venda do veículo.