Machine Learning Capstone Project

Nanodegree Engenheiro de Machine Learning

Marcus Vinicius de Oliveira Cruz
22 de Fevereiro de 2018

Time Series Forescasting - Shanghai license plate buildding price prediction

1. Project Overview

Como conclusão do nanodegree Engenheiro de Machine Learning e por uma possibilidade de uma consultoria de machine learning resolvi ir a fundo em um modelo de time series forecasting. Um modelo de série temporal basicamente consiste em um modelo estatístico que analisa uma variação temporal e consegue realizar previsões.

Escolhi o dataset Shanghai license plate bidding price prediction para construir o meu modelo do capstone. Esse é um dataset que pertence ao Kaggle.¶

O aumento da propriedade e uso de automóveis na China nas últimas duas décadas aumentou o consumo de energia, piora a poluição do ar e congestionamento exacerbado. O governo de

Xangai adotou um sistema de leilão para limitar o número de placas emitidas para cada mês. O conjunto de dados contém dados históricos de leilões de janeiro de 2002 a outubro de 2017.

como funciona o sistema de leilão: um preço inicial é dado no início do leilão, os licitantes só podem oferecer até 3 vezes por cada leilão e só podem marcar para cima ou para baixo dentro de 300 CNY (aproximadamente 46 USD) por cada lance. No final de cada leilão, apenas o n superior (número de placas que serão emitidas para o mês) receberá as placas de licença ao custo de suas propostas. A oferta n. ° será o preço mais baixo do mês. Por favor, note que os leilões são realizados on-line e cada licitante não poderá ver outros lances.

Colunas:

Data: janeiro de 2002 a outubro de 2017 (observe que faltam em fevereiro de 2008)

- num_bidder *: número de cidadãos que participam do leilão para o mês
- num_plates *: número de placas que serão emitidas pelo governo para o mês
- low_deal_price *: explicado acima, em CNY
- avg_deal_price *: preço médio do negócio, no CNY (observe que, como cada lance só pode ser marcado para cima ou para baixo no prazo de 300, não está se afastando muito do preço mais baixo)

O objetivo é prever o preço low*deal* para cada mês, o resultado real será atualizado no final de cada mês

o conjunto de dados é raspado de http://www.51chepai.com.cn/paizhaojiage/

Contato: ran_su147@hotmail.com

Contexto

Xangai usa um sistema de leilões para vender um número limitado de matadouros para compradores de automóveis com combustível fóssil todos os meses. O preço médio desta placa de licença é de cerca de US $ 13.000 e muitas vezes é referido como "a peça de metal mais cara do mundo". Então, nosso objetivo é prever o preço médio ou o preço mais baixo para o próximo mês. Este conjunto de dados será atualizado mensalmente constantemente.
A placa de matrícula de Xangai foi aclamada como "a pele de metal mais cara do mundo", a licença de carro privada de Xangai desde o início do leilão em 1986, com mais de 10 anos de desenvolvimento contínuo, sistema de licença de carro privado desde o início do leilão de alto preço, Xangai O processo de leilão de matrículas continuou mudando, até o atual leilão inestimável.

Sem preocupações, placa de placas de rede, retrato, lança carta de preços da matrícula de Xangai, para facilitar você a verificar e comparar o preço da placa de matrícula 2002-2017 Xangai.
Como uma cidade internacional, Xangai tornou-se o centro de atenção do Leste Asiático, juntamente com o desenvolvimento acelerado, o tráfego tornou-se um dos problemas mais pesados. Vamos tentar fazer uma análise sobre o preço geral da licença do carro.

Os principais objetivos das séries temporais consistem em:

- Compreender o mecanismo gerador da série;
- Predizer o comportamento futuro da série.

Dessa forma ao entender um mecanismo de série temporal pode-se:

- Descrever e analisar de fato o comportamento da série temporal;
- Entender sobre as periodicidades presentes nas séries temporais;
- Entender o que o ocasiona o comportamento da série temporal;
- Controlar a trajetória da série temporal.

Dessa forma para explicar o modelo vamos utilizar 4 aplicações de funções de time series forecasting, três delas focadas em passagens de avião é um deles focado em venda de tratores. Dessa forma é possível observar o comportamento dos algoritmos de forecasting em dados de testes validados.

2. Problem Statement (declaração do problema)

O maior problema em questão é conseguir quantificar uma série temporal e saber qual a melhor forma de tratar tais dados. Para esse projeto em específico o problema e implementar um modelo de forecasting em um dataset com registros de valores de passagens de avião.

Teoricamente um modelo de forecasting consiste na visualização de dados acerca da variação da série temporal. Em um determinado momento, após o algoritmo ter entendido sobre a variação de preços e com as implementações de condições já pré estabelecidas e conhecidas acontece uma previsão.

Implementar um modelo de forecasting é basicamente um problema de aprendizagem supervisiona, e pode ser entendimento como uma regressão linear. Visto que o algoritmo vai fazer vasculhar a base de dados em busca de padrões de irregularidades afim de implementar previsões assertivas sobre o comportamento futuro dos dados.

3. Metrics

Pelo fato de o forecasting usar muito a visualização dos dados o forecasting poderá ser visto funcionando junto a essas visualizações. Dessa forma deverão sempre ser feitos testes baseados no comportamento real dos dados.

Como métrica de avaliação principal usei a comparação entre os algoritmos implementados.

Analysis

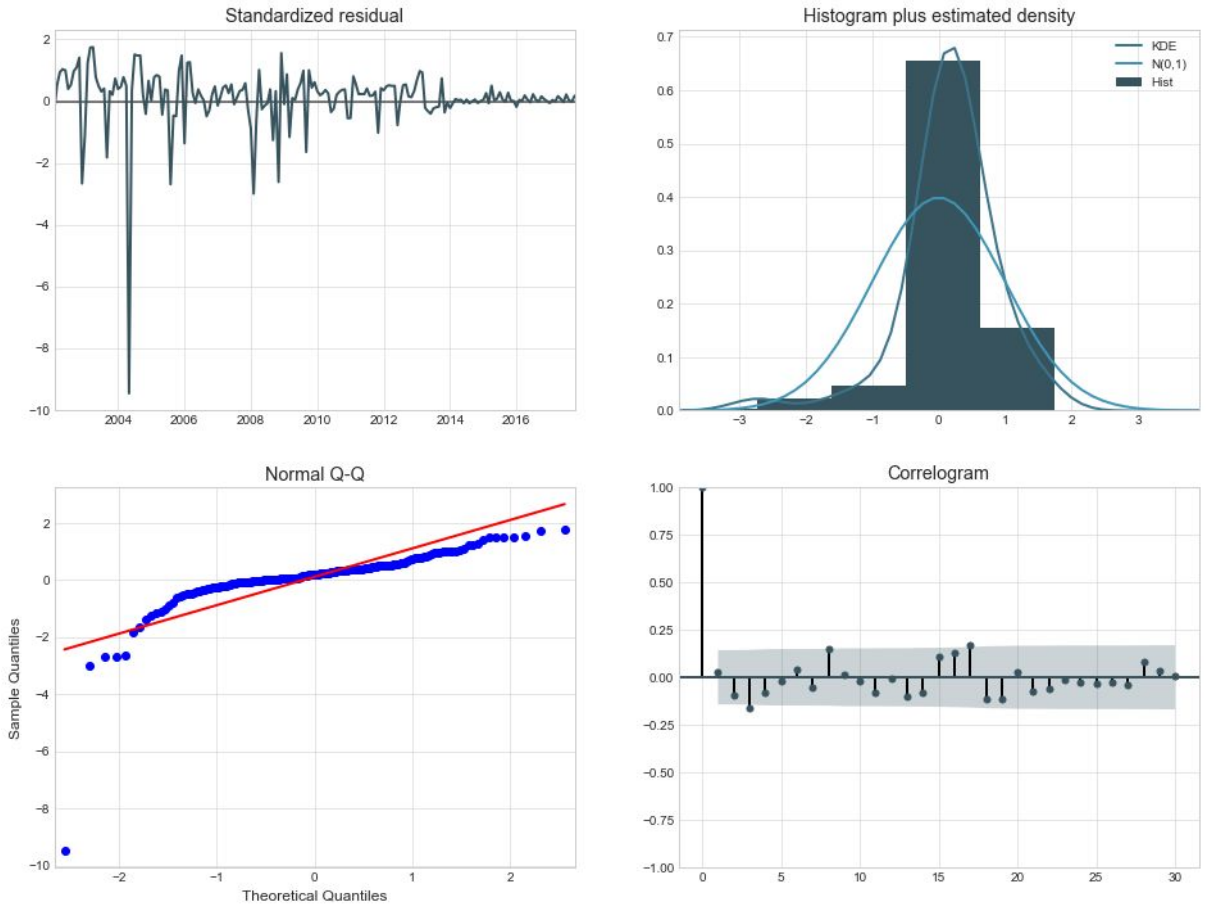
4. Data Exploration

Durante a análise exploratória dos dados várias questões foram sendo respondidas. Pelas bibliotecas de time series forecasting usarem muito de visualizações para mostrar resultados nos gráficos.

5. Exploration Visualization

Foi fornecida uma visualização que resume ou extrai uma característica ou característica relevante sobre o conjunto de dados ou os dados de entrada com uma discussão aprofundada. As pistas visuais estão claramente definidas.

O gráfico abaixo mostra um plot com um diagnóstico sobre as melhores visualizações para representar o modelo.



6. **Algorithms and Techniques**

Algoritmos e técnicas utilizadas no projeto são cuidadosamente discutidas e devidamente justificadas com base nas características do problema.

Uma série temporal basicamente consiste em na coleta de dados em intervalos de tempos constante. De forma que sejam criadas visualizações de dados para apontar hipóteses mais conclusivas durante a análise de dados.

Time series forecasting é um modelo de aprendizagem supervisionada, como tal precisa de exemplos claros e concisos para treinamento do modelo em questão.

- Análise Exploratória dos Dados
- Regressão Linear
- Support Vector Machines (SVM)
- Principal Component Analysis - PCA
- K Nearest Neighbors
- Árvores de decisão e florestas aleatórias
- Time series forecasting

Methodology

7. **Data Preprocessing**

Na etapa de pré processamento dos dados apliquei um padrão na análise de dados que consiste em:

1. Reunir os dados.
2. Avaliar os dados.
3. Limpar os dados

8. **Implementation**

O processo para o qual métricas, algoritmos e técnicas foram implementadas com os conjuntos de dados ou dados de entrada dados foi completamente documentado. As complicações ocorridas durante o processo de codificação são discutidas.

Análise exploratória dos dados

A partir do gráfico, podemos ver o preço da placa aumentou drasticamente, especialmente a partir de 2014. Ele subiu de cerca de 20.000 yuan para quase 90.000 yuan com em 16 anos. A

característica de Anther que podemos ver a partir do gráfico é o preço de oferta mais baixo é muito próximo do preço médio da oferta, e a diferença está cada vez mais próxima e especialmente após 2013, a diferença está dentro de 200 yuan.

A segunda parcela mostrou que os participantes do lance de licença apresentaram uma taxa de crescimento ultrajante, em particular acontecem em torno de 2014. No entanto, o número de ofertas de placas para público é significativamente menor que o número do lance. O número de lances passou de média de 4.373 por ano para 253.335 Aplicações por ano. Sem licença, a placa de licença só oferece 2,654 a 11,023 por ano, em média.

Análise do preço médio da placa de licença

Agora vamos investigar mais sobre os dados, comparando-os no ano. Vamos agrupar os dados por ano, em seguida, tomar o preço médio de cada ano. Nós mostramos pela primeira vez um gráfico de barras da relação entre os candidatos da placa de automóveis e o número de placas de automóveis. O gráfico demonstrado a partir de 2014, o número de candidatos crescem excessivamente fora das placas. Em seguida, o gráfico da taxa de crescimento mostrou que o número de placas de licença são relativamente estáveis, compare o número de candidatos. A taxa de cobrança de placa de carro apoiou a conclusão anterior, antes de 2014, a taxa vencedora é de 30% ou mais. A partir de 2014, a taxa vencedora diminuiu para menos de 10%, hoje em dia é de cerca de 5%. Além disso, o crescimento do preço médio está dentro de 10% a partir de 2015. Então vamos traçar a taxa de crescimento com 2002 como linha de base, veja como ela muda a cada ano em relação ao ano de 2002. Podemos ver, a taxa vencedora, o preço médio é gradualmente converge para o estável a partir de 2015. Eu acho que o principal motivo é que o mercado está gradualmente saturado.

Sabemos que o preço do negócio médio é muito próximo do preço mais baixo. Então, estamos tentando usar o número de licitadores para prever o preço mais baixo. Podemos ver quando o número do lance> 50,000, pode haver uma relação linear. Nosso modelo apoiou nossa conjectura, o valor p menor do que 0,05. Então, se conhecemos o número total de licitantes, então podemos calcular o preço mais baixo, e pelo preço mais baixo, adicionamos 300-500Yuan, será na Loteria.

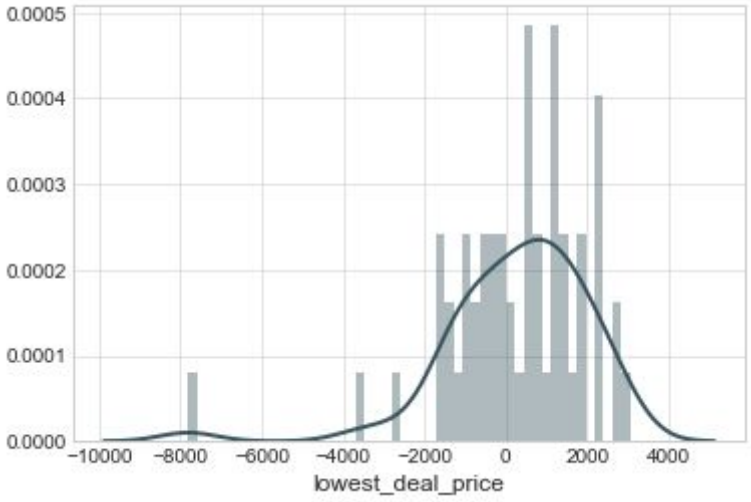
9. Refinement

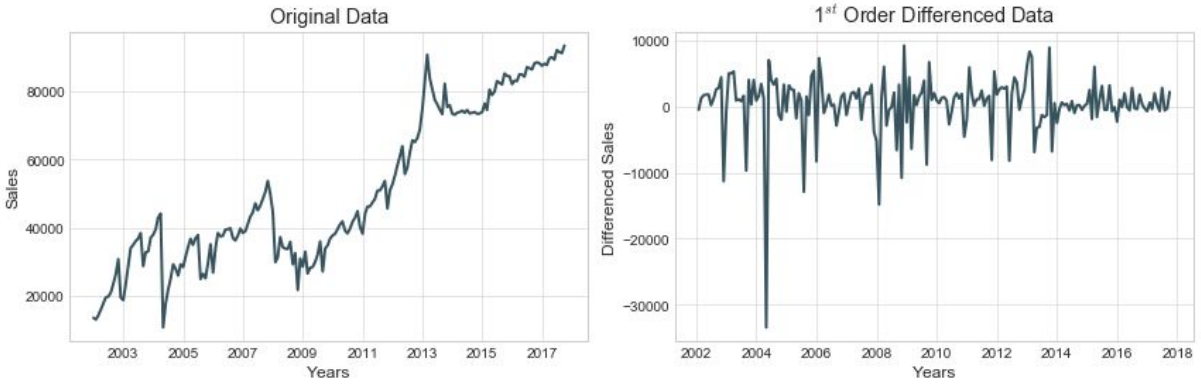
O processo de melhoria nos algoritmos e técnicas utilizadas está claramente documentado. As soluções iniciais e finais são relatadas, juntamente com soluções intermediárias, se necessário.

Results

10. Model Evaluation and Validation

As qualidades do modelo final - como os parâmetros - são avaliadas em detalhes. Algum tipo de análise é usado para validar a robustez da solução do modelo.





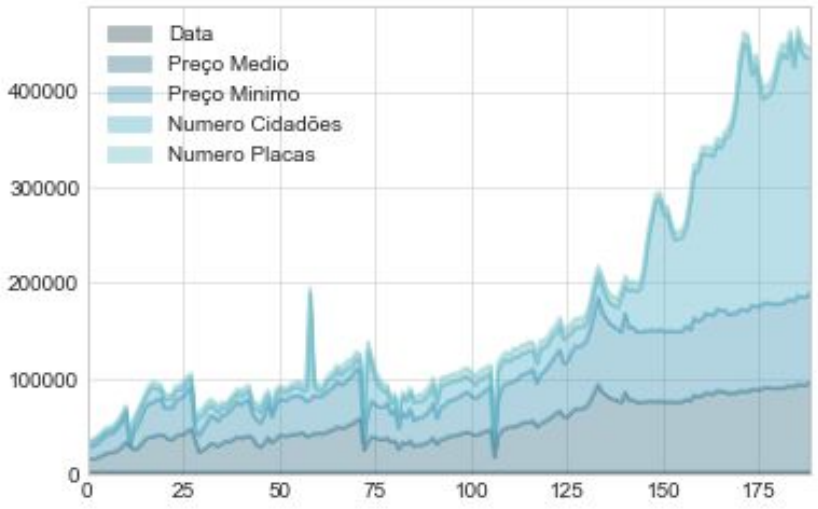
11. Justification

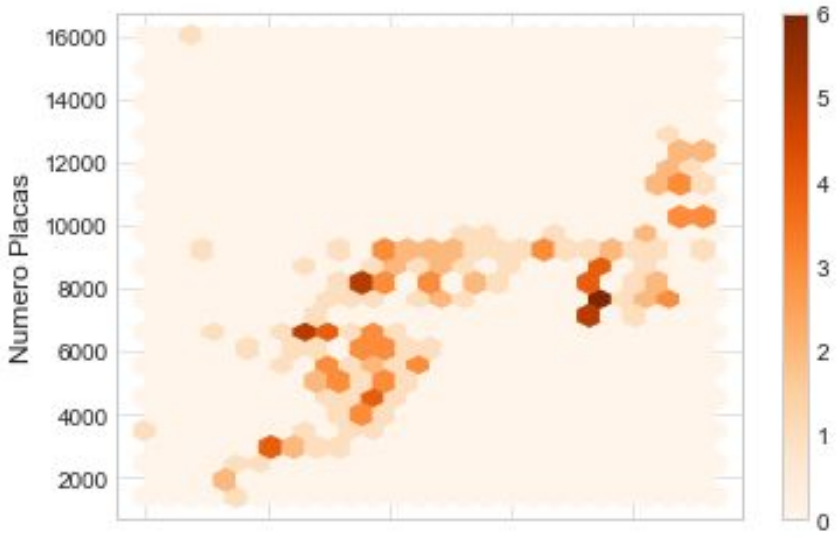
Os resultados finais são comparados com o resultado ou o limite de referência com algum tipo de análise estatística. A justificação é feita para saber se o modelo final e a solução são suficientemente significativos para ter resolvido adequadamente o problema.

Conclusion

12. Free-Form Visualization

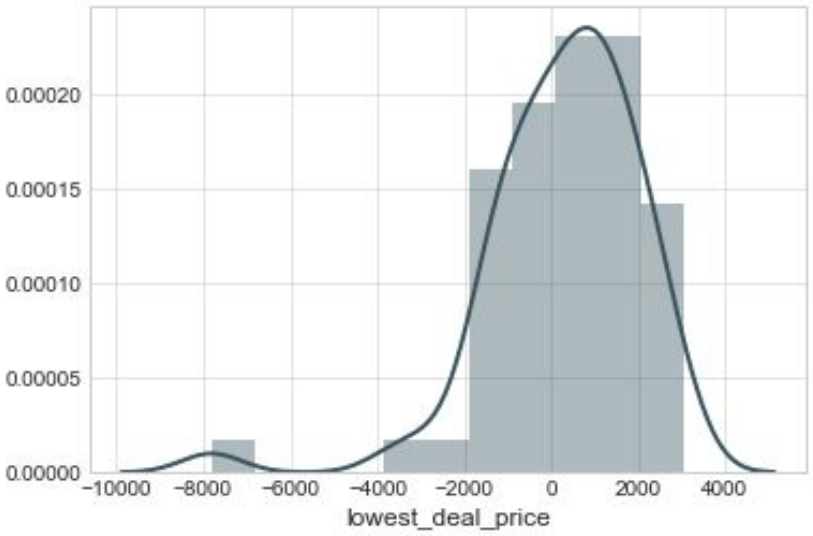
Foi fornecida uma visualização que enfatiza uma qualidade importante sobre o projeto com uma discussão aprofundada. As pistas visuais estão claramente definidas.

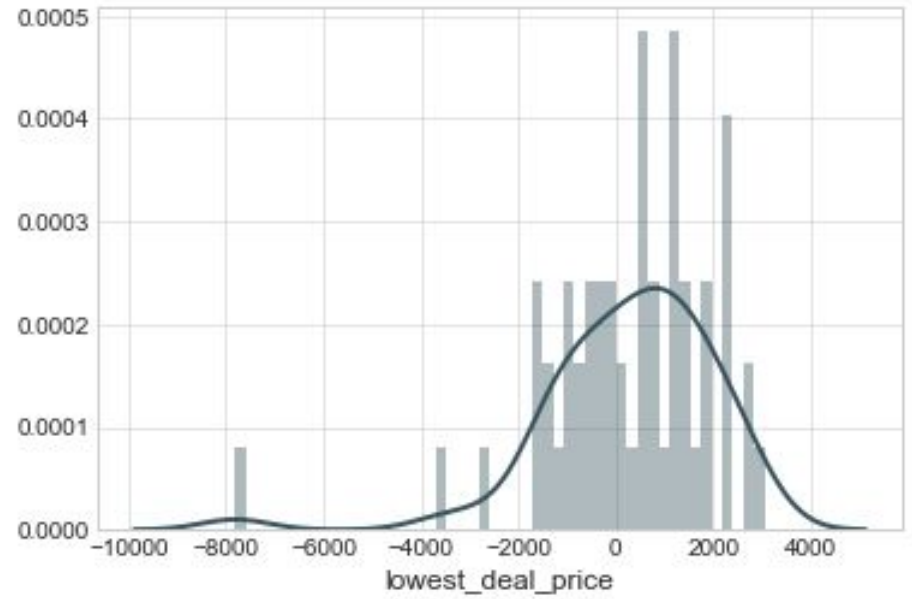


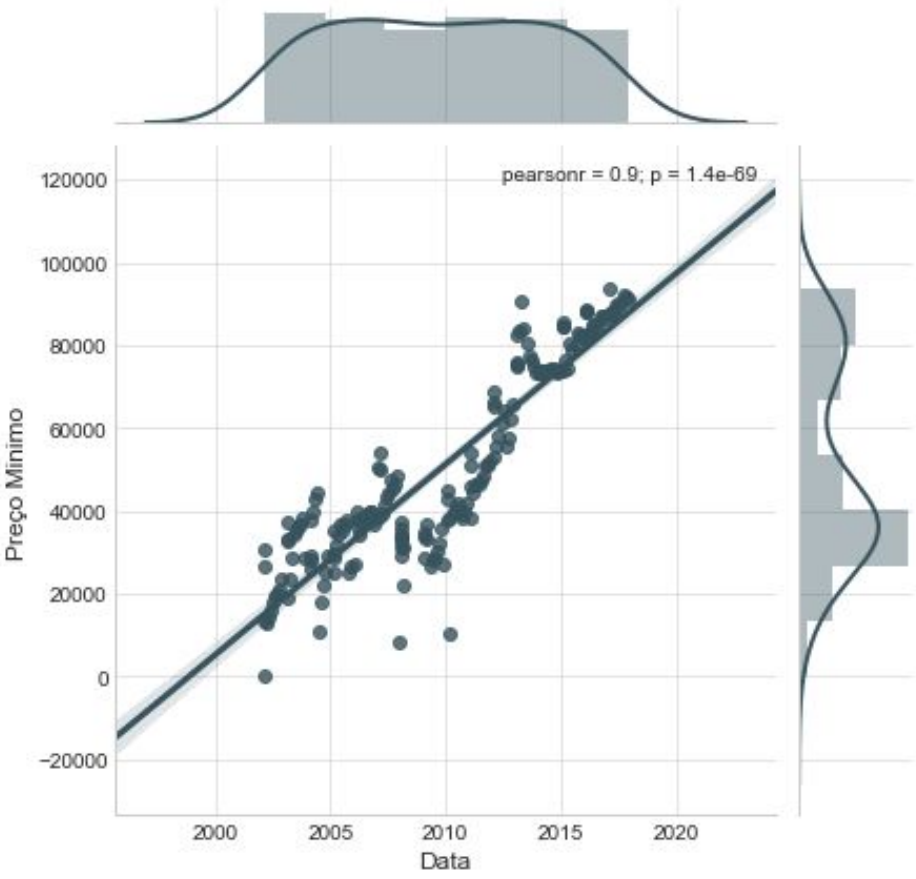


13. Reflection

O aluno resuma de forma adequada a solução de problemas de ponta a ponta e discute um ou dois aspectos específicos do projeto que eles acham interessante ou difícil.

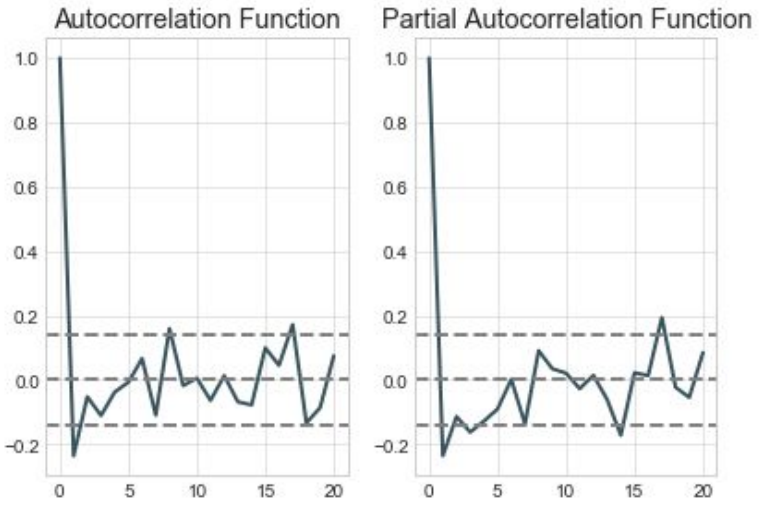


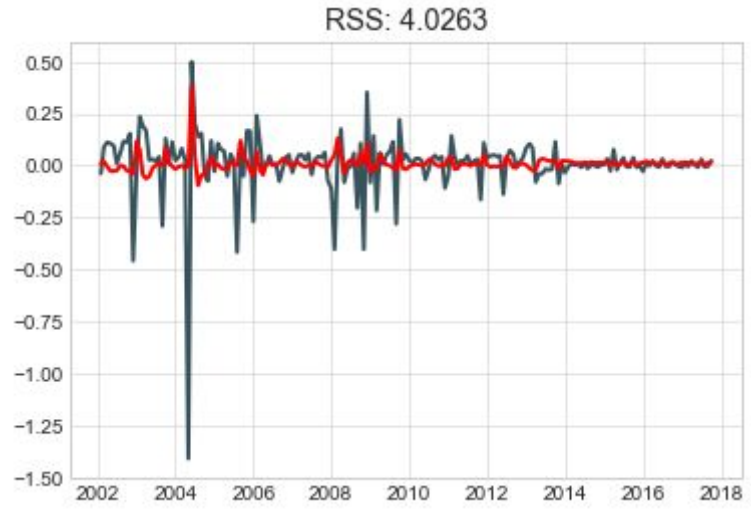


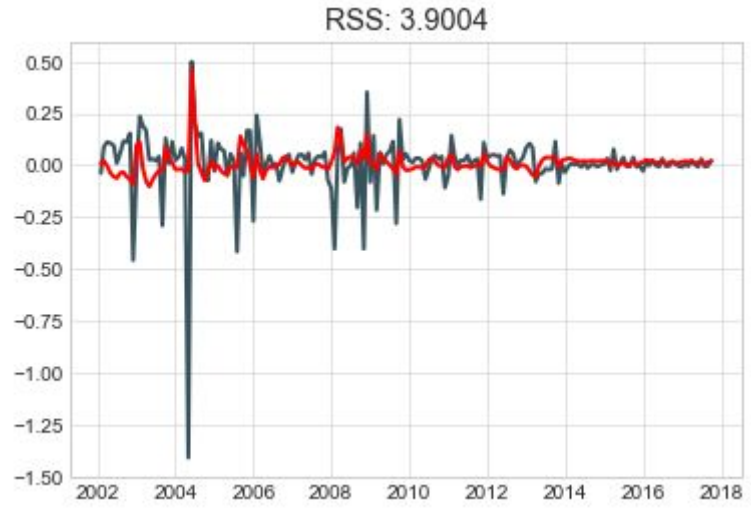


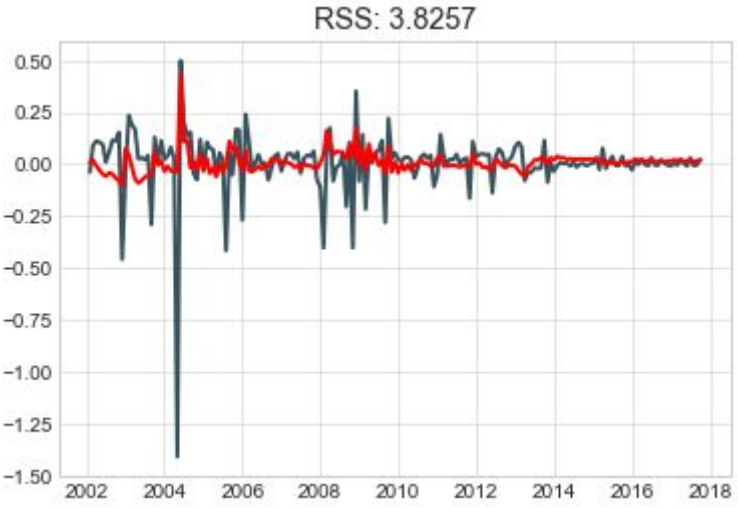
14. Improvement

É feita uma discussão sobre como um aspecto da implementação poderia ser melhorado. As soluções potenciais resultantes dessas melhorias são consideradas e comparadas / contrastadas com a solução atual.







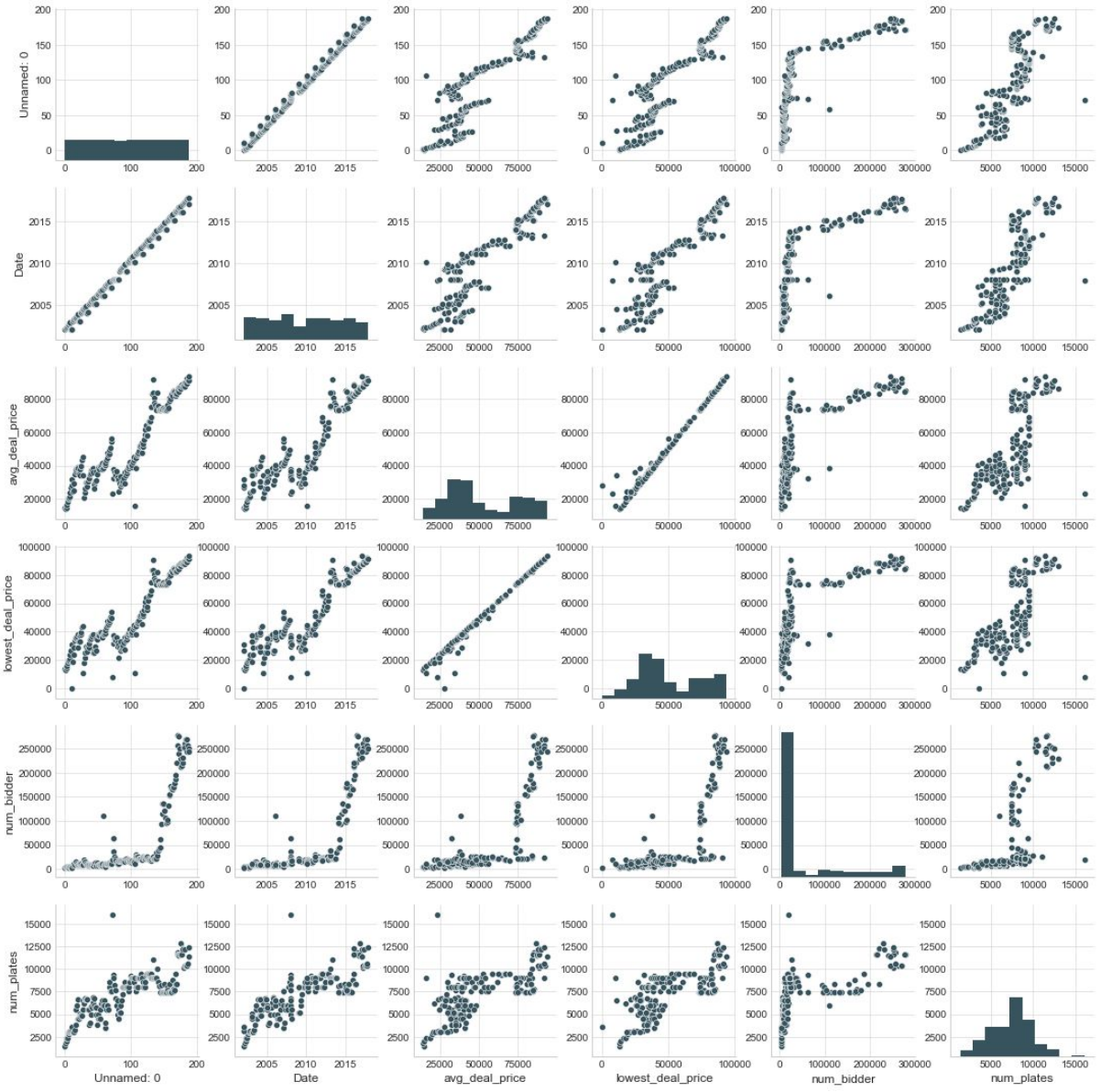


Quality

15. Presentation

Em todo o projeto procurei deixar os textos o mais claros e explicativos possíveis. Revisei o projeto diversas vezes e acredito que os erros de português serão mínimos.

Dei a devida referência a todo material que consultei, consultei não só artigos de internet, mas também artigos acadêmicos. O referencial teórico está gigante pois procurei entender de fato os problemas e buscar a melhor ferramenta para resolver os problemas iniciais.

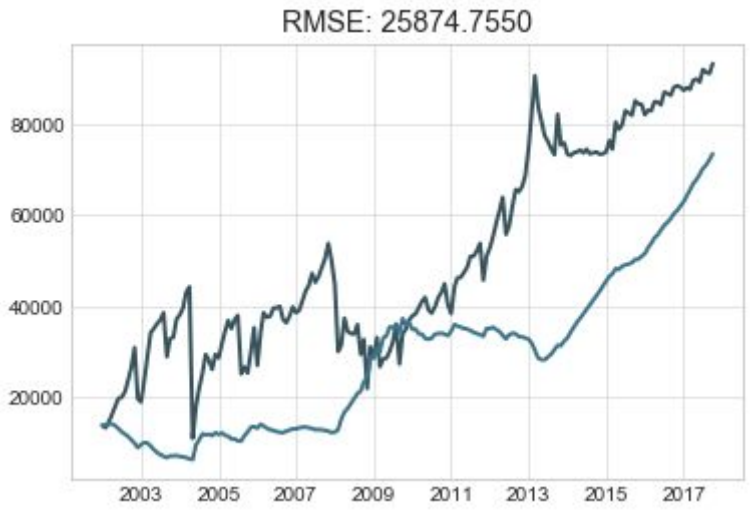


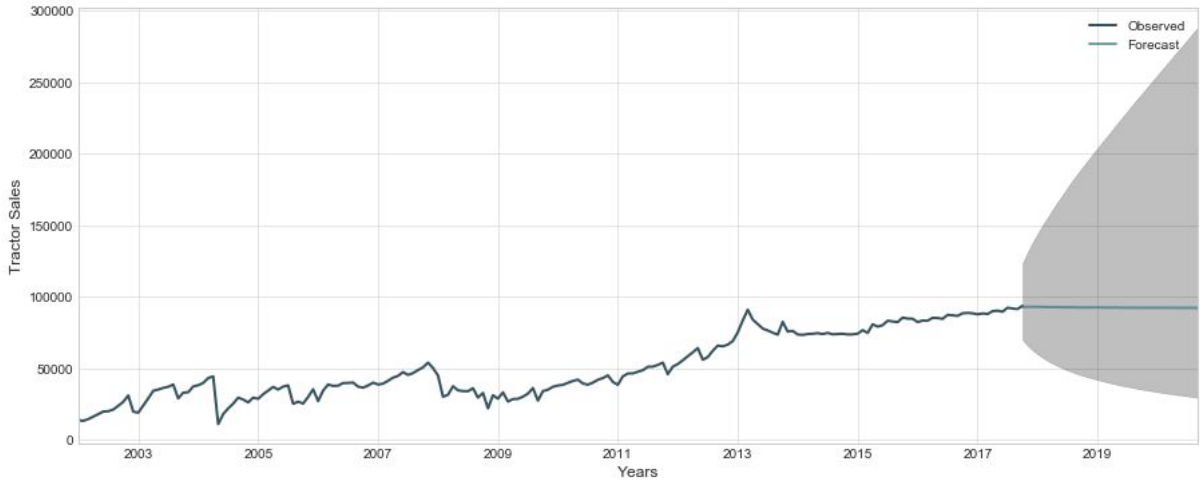
16. Functionality

Conforme orientação da Udacity o Jupyter Notebook está completamente explicativo. Procurei documentar todo o processo de criação, assim como toda a validação das hipóteses do início ao final do projeto.

Esse modelo de time series forecasting está funcional e conseguiu atingir os resultados esperados quando ainda estava apenas especulando sobre os resultados do projeto.

Como eu foquei em deixar um projeto interessante não somente para uma aplicação específica acredito que esse projeto está atendendo perfeitamente aos pré requisitos.





17. Referências

https://www.kaggle.com/bogof666/shanghai-car-license-plate-auction-price/data

https://www.kaggle.com/bazingasu/shanghai-license-plate-bidding-price-prediction/data

http://deeplearningbook.com.br/capitulo-9-a-arquitetura-das-redes-neurais/

https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/

https://codeburst.io/jupyter-notebook-tricks-for-data-science-that-enhance-your-efficiency-95f98d3adee4

https://towardsdatascience.com/why-you-should-forget-for-loop-for-data-science-code-and-embrace-vectorization-696632622d5f

https://towardsdatascience.com/10-machine-learning-algorithms-you-need-to-know-77fb0055fe0

http://www.cbcity.de/timeseries-decomposition-in-python-with-statsmodels-and-pandas

https://bibliotecadigital.ipb.pt/bitstream/10198/12709/1/Artur%20Jorge%20Ferreira%20da%20Costa%20Dias.pdf

http://www.redalyc.org/pdf/3291/329147536007.pdf

http://www.ceel.eletrica.ufu.br/artigos/ceel2016_artigo094_r01.pdf

https://repositorio.bc.ufg.br/tede/bitstream/tede/7563/5/Disserta%C3%A7%C3%A3o%20-%20Ricardo%20Henrique%20Fonseca%20Alves%20-%202017.pdf

http://ftp.cptec.inpe.br/labren/publ/teses/DISSERTACAO_RICARDO-GUARNIERI.pdf

http://www.confea.org.br/media/contecc2017/eletrica/1_audrnanpdrsg.pdf

http://www.inovarse.org/sites/default/files/T14_0291_5.pdf

http://www.redalyc.org/html/3291/329147536007/

https://repositorio.ufu.br/handle/123456789/14569

https://repositorio.ufsc.br/bitstream/handle/123456789/178026/TCC_Final_Jhuan_Souza.pdf?sequence=1&isAllowed=y

https://www.producaoonline.org.br/rpo/article/view/2542/1596

revistas.ufpr.br/rber/article/download/48431/pdf

https://pt.stackoverflow.com/questions/192098/como-funciona-uma-rede-neural-artificial

https://martin-thoma.com/classification-with-pybrain/

http://conteudo.icmc.usp.br/pessoas/andre/research/neural/

http://www.din.uem.br/ia/neurais/

http://www.cerebromente.org.br/n05/tecnologia/rna.htm

ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia006_03/topico5_03.pdf

http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0016231_04_cap_05.pdf

https://www.embarcados.com.br/redes-neurais-artificiais-introducao/

https://periodicos.utfpr.edu.br/recit/article/view/4330/Leandro

http://www2.ica.ele.puc-rio.br/Downloads/33/ICA-introdu%C3%A7%C3%A3o%20RNs.pdf

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1982-21702017000100150&lng=pt&tlng=pt

http://www.scielo.br/scielo.php?pid=S1678-86212017000300103&script=sci_arttext

https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

https://www.monolitonimbus.com.br/processos-estacionarios/

https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

http://www.portalaction.com.br/series-temporais/11-estacionariedade

http://www.icad.puc-rio.br/cfeijo/pdf/revis%C3%A3o%20b%C3%A1sica%20s%C3%A9ries%20temporais_material%20de%20apoio_curso%20teoria%20macroeconomica_PPGE%20UFF.pdf

https://www.ime.unicamp.br/~hlachos/MaterialSeries.pdf

http://www.inf.ufsc.br/~marcelo.menezes.reis/Cap4.pdf

https://www.researchgate.net/publication/229040330_JTIMESAT_uma_ferramenta_para_a_visualizacao_de_series_temporais_de_imagens_de_satelite

http://bdm.unb.br/bitstream/10483/7239/1/2013_JoseRobertoGoncalvesdeRezendeFilho.pdf

https://www.maxwell.vrac.puc-rio.br/16824/16824_4.PDF

https://www.maxwell.vrac.puc-rio.br/24787/24787_4.PDF

http://conteudo.icmc.usp.br/pessoas/ehlers/stemp/stemp.pdf

http://cdsid.org.br/sbpo2015/wp-content/uploads/2015/08/140250.pdf

https://www.marinha.mil.br/spolm/sites/www.marinha.mil.br.spolm/files/101711_0.pdf

https://www.lume.ufrgs.br/bitstream/handle/10183/31034/000782115.pdf?sequence=1

http://www2.ufersa.edu.br/portal/view/uploads/setores/232/TCC%20-%20VALCIANO%20CAMILO%20GURGEL.pdf

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1982-21702017000100150&lng=pt&tlng=pt

http://www.ime.eb.br/arquivos/teses/se4/mec2008/2008Bianca.pdf

http://repositorio.unicamp.br/bitstream/REPOSIP/267746/1/Conti_JoseCarlos_M.pdf

http://www.ctec.ufal.br/professor/cfs/Sul_Sud06%20-%20Series.pdf

http://pdf.blucher.com.br.s3-sa-east-1.amazonaws.com/marineengineeringproceedings/spolm2015/140011.pdf

http://www.portalaction.com.br/series-temporais/15-modelos-para-series-temporais

http://www.scielo.br/scielo.php?pid=S1678-86212017000300103&script=sci_arttext

https://www.researchgate.net/publication/289479535_Previsao_de_energia_eletrica_modelagem_e_uso_de_combinacoes_de_previsoes

https://www.ufrgs.br/sbai17/papers/paper_506.pdf

http://www.scielo.org.co/pdf/eia/n26/en_n26a09.pdf

http://www.sciencedirect.com/science/article/pii/S1877050915015641

http://www.uff.br/engevista/seer/index.php/engevista/article/viewFile/433/236

http://www.ufjf.br/pgmc/files/2011/05/Disserta%C3%A7%C3%A3o-Guilherme-G-Neto-18-08.pdf

http://www.exatas.ufpr.br/portal/degraf_paulo/wp-content/uploads/sites/4/2014/09/EE022-08-08.pdf

http://www.datascienceinstitute.com.br/forecast-de-consumo-de-energia-eletrica/

https://docs.microsoft.com/pt-br/azure/machine-learning/preview/scenario-time-series-forecasting

https://translate.google.com.br/translate?sl=en&tl=pt&js=y&prev=_t&hl=pt-BR&ie=UTF-8&u=http%3A%2F%2Fwww.scielo.br%2Fscielo.php%3Fscript%3Dsci_arttext%26pid%3DS1678-86212017000300103%26lng%3Dpt%26tlng%3Dpt&edit-text=

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212017000300103&lng=pt&tlng=pt

http://acervo.ufvjm.edu.br/jspui/bitstream/1/1327/1/rodrigo_magalhaes_mota_santos.pdf

http://tede2.pucgoias.edu.br:8080/bitstream/tede/2484/1/Paulo%20Henrique%20Borba%20Florencio.pdf

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.4455&rep=rep1&type=pdf

https://gab41.lab41.org/the-10-algorithms-machine-learning-engineers-need-to-know-f4bb63f5b2fa

http://minerandodados.com.br/index.php/2017/05/19/prevendo-precos-de-acoes-da-bolsa-de-valores-com-machine-learning/

https://fga.unb.br/articles/0000/5556/TCC_Hialo_Muniz.pdf

http://www.feis.unesp.br/Home/departamentos/engenhariaeletrica/pos-graduacao/327-dissertacao_ciceromarcelo.pdf

https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

http://www.semantix.com.br/10-algoritmos-de-machine-learning/

http://www.scielo.br/scielo.php?pid=S1678-86212017000300103&script=sci_arttext

https://fga.unb.br/articles/0000/7804/TCC_Hialo_Muniz.pdf

http://www.leec.eco.br/downloads/R-tutorial-de-bolso.pdf