**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Leandro da Silva Câmara**

**PREVISÃO DE VENDAS COM PROPHET**

Belo Horizonte

2020

**Leandro da Silva Câmara**

**PREVISÃO DE VENDAS COM PROPHET**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte

2020

**SUMÁRIO**

[**1. Introdução**](#_i9f8fyqc8w70) **4**

[**1.1. Contextualização**](#_41eq00gw0695) **4**

[**1.2. O problema proposto**](#_h2egnqa93wmm) **4**

[**2. Coleta de Dados**](#_7tdpo1gkxe7p) **5**

[**3. Processamento/Tratamento de Dados**](#_3l80a95igbsm) **7**

[**3.1. Pré-Processamento - Análise e Exploração de Dados**](#_t9azg7jq4bhi) **7**

[**3.2. Pré-Processamento - Machine Learning**](#_3xt2g1wb77xd) **8**

[**4. Análise e Exploração dos Dados**](#_v0mxjarokkj) **9**

[**5. Criação de Modelos de Machine Learning**](#_ir0p3vhc3em6) **11**

[**6. Apresentação dos Resultados**](#_8fningsm3psy) **11**

[**7. Links**](#_bguw696ycgsv) **12**

[**REFERÊNCIAS**](#_muecavpqtr4b) **13**

# 1. Introdução

## 1.1. Contextualização

Com o crescimento da adoção de novas tecnologias e o aumento na competitividade por parte das empresas, ter informações relevantes que auxiliem nas tomadas de decisões, pode ser um diferencial significativo em relação aos demais concorrentes.

Dentre diversas informações que podem auxiliar os gestores, o acesso às previsões de vendas pode ajudar em mudanças e ações estratégicas com antecedência e potencializar os lucros ou minimizar os efeitos negativos no negócio, em casos de quedas das vendas e/ou receita.

## 1.2. O problema proposto

A previsão de vendas pode embasar as decisões estratégicas em diferentes áreas, desde o volume de matéria-prima a ser comprado até às contratações e demissões de funcionários, ou, ainda, às decisões de investimento em infraestrutura.

Ao conhecer o que o futuro reserva para a empresa, fica muito mais fácil prever e prevenir falhas ou crises. Além disso, a empresa pode se organizar de forma a evitar que esforços sejam desperdiçados.

Durante a projeção de vendas, também são levantadas novas oportunidades e potenciais ameaças, o que garante à empresa tempo hábil para planejar alternativas de aproveitar o que foi identificado em seu benefício.

Além disso, se a empresa entende o mercado de atuação e projeta suas expectativas de vendas, consegue controlar melhor os custos, os investimentos, as pessoas e o estoque, evitando perdas e prejuízos.

Visando os benefícios proporcionados pela previsão de vendas, será analisado um conjunto de dados, disponibilizado (por meio da plataforma *Kaggle*) por uma das maiores empresas russas de software – *1C Company*, que consiste em vendas diárias de um determinado período.

Por meio da exploração dos dados, propõe-se identificar informações que possam ser relevantes para a empresa, como por exemplo, lojas e categorias de produtos relacionadas com as maiores e menores vendas e geração de receita. Além disso, será realizado análises para verificação da existência de tendências e sazonalidades.

O período analisado refere-se aos meses entre Janeiro de 2013 a Outubro de 2015, totalizando 34 meses. Com essas informações, pretende-se elaborar um modelo, utilizando a biblioteca para construção de séries temporais do *Facebook* – denominada *Prophet*, que seja capaz de prever os próximos 5 meses de vendas e receita da empresa.

# 

# 2. Coleta de Dados

Os dados referente às vendas diárias da empresa *1C Company* foram obtidos no dia 14 de Janeiro de 2020, por meio de arquivos em formato CSV, disponibilizados na plataforma *Kaggle[[1]](#footnote-0)*. O conjunto de dados (*dataset*) abrange as vendas realizadas entre Janeiro de 2013 e Outubro 2015. O *dataset* é composto por **2.935.849 registros** de vendas e 3 arquivos, sendo estes:



* ***sales.csv*** - Dados históricos (vendas diárias) de Janeiro de 2013 a Outubro de 2015;

|  |  |  |
| --- | --- | --- |
| **Nome da coluna** | **Descrição** | **Tipo** |
| *date* | Data no formato “*dd.mm.YYYY*”. | *String* |
| *date\_block\_num* | Número sequencial dos meses entre Janeiro de 2013 (0) e Outubro de 2015 (34). | *Integer* |
| *shop\_id* | Identificador da loja. | *Integer* |
| *item\_id* | Identificador do produto. | *Integer* |
| *item\_price* | Preço do produto. | *Float* |
| *item\_cnt\_day* | Quantidade de produtos vendidos. | *Integer* |

* ***items.csv*** - Informações complementares sobre os itens/produtos;

|  |  |  |
| --- | --- | --- |
| **Nome da coluna** | **Descrição** | **Tipo** |
| *item\_id* | Identificador do produto. | *Integer* |
| *item\_name* | Nome do produto. | *String* |
| *item\_category\_id* | Identificador da categoria do produto. | *Integer* |

* ***item\_categories.csv*** - Informações complementares sobre as categorias de itens.

|  |  |  |
| --- | --- | --- |
| **Nome da coluna** | **Descrição** | **Tipo** |
| *item\_category\_id* | Identificador da categoria do produto. | *Integer* |
| *item\_category\_name* | Nome da categoria do produto. | *String* |

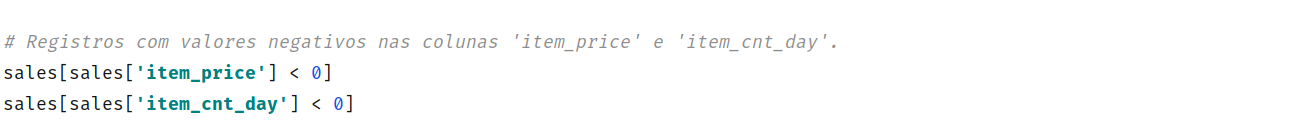
# 3. Processamento/Tratamento de Dados

Durante o processo de análise/exploração dos dados e construção do modelo de previsão de vendas, foi fundamental a realização de alguns processamentos/tratamentos de dados.

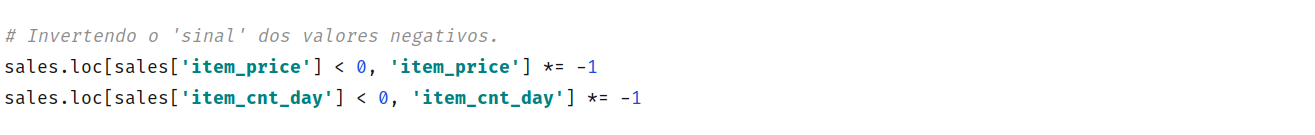
## 

## 3.1. Pré-Processamento - Análise e Exploração de Dados

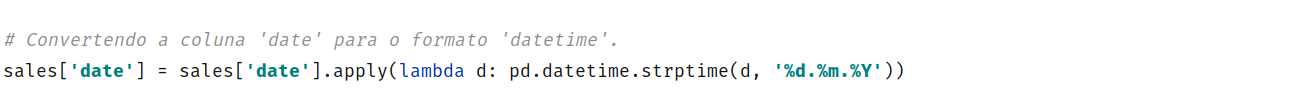
Um dado importante, mas ausente no *dataset*, é o valor total da venda (diária). Portanto, uma nova coluna foi criada - denominada “*total\_value*”, por meio da multiplicação das colunas “*item\_price*” e “*item\_cnt\_day*”. Entretanto, foram identificados **7357 registros** com valores inválidos (negativos) em tais colunas.



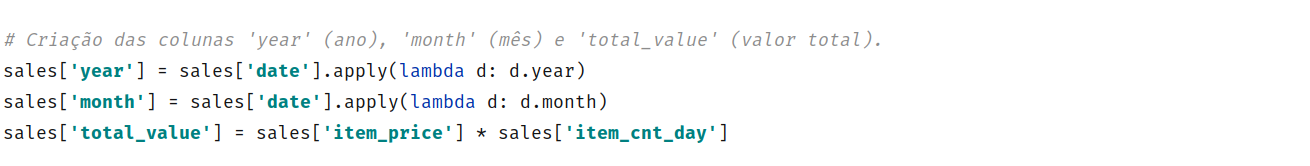
Levando em consideração a premissa de que todas as vendas estão relacionadas com pelo menos um produto, todas as vendas com quantidades e preços negativos de produtos foram transformados para valores positivos.



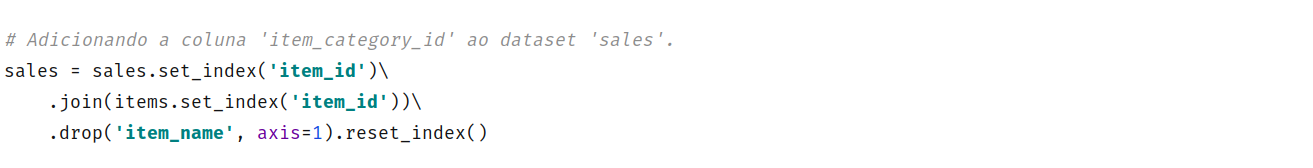
Devido ao fato da coluna “*date*” ser do tipo “*string*”, fez-se necessário a conversão dos dados provenientes de tal coluna para o formato “*datetime*”, pois neste novo formato, torna-se possível análises com base em meses e anos.



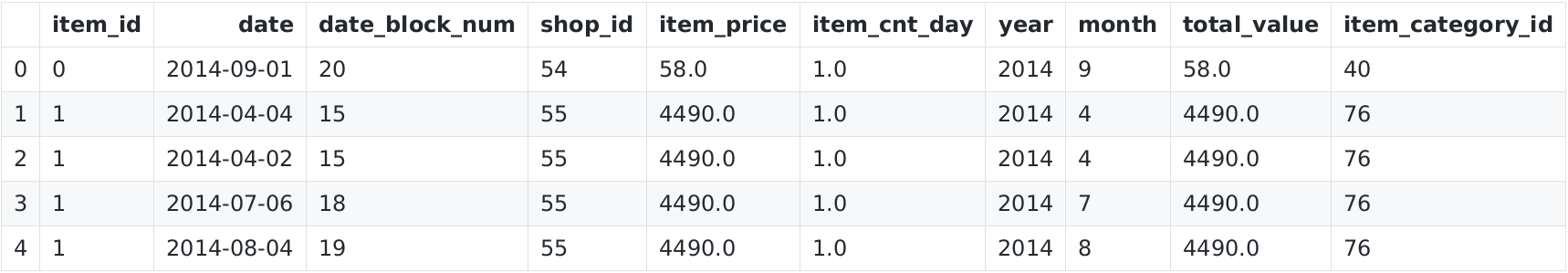
Após o conversão realizada na coluna “*date*”, foi possível adicionar à cada registro as colunas “*year*” (ano) e “*month*” (mês). Além disso, foi incluído a coluna “*total\_value*” (valor total).



Com o objetivo de realizar análises através do relacionamento das vendas com as categorias de produtos, foi incluído ao *dataset* “*sales*” a coluna “*item\_category\_id*”, mediante à junção dos *datasets* “*sales*” e “*items*”, por meio da coluna “*item\_id*”.

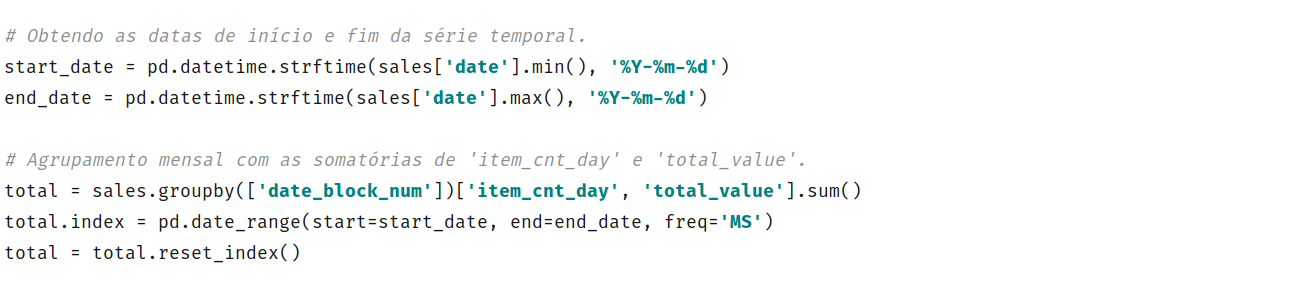


[...]



## 3.2. Pré-Processamento - *Machine Learning*

Para criação do modelo de *Machine Learning*, foi necessário um pré-processamento no *dataset* “*sales*”, para que as vendas fossem agrupadas mensalmente, de modo a obter o somatório das vendas e da receita.



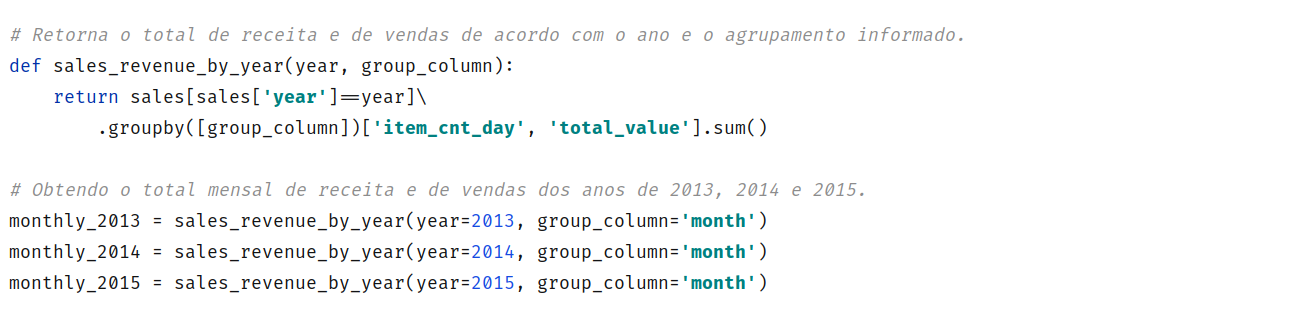
Após o agrupamento mensal, foi necessário dividir resultado, de modo que o total das vendas ficasse separada da receita. Esta partição é fundamental, pois será criado um modelo (Machine Learning) para cada *dataset* gerado.



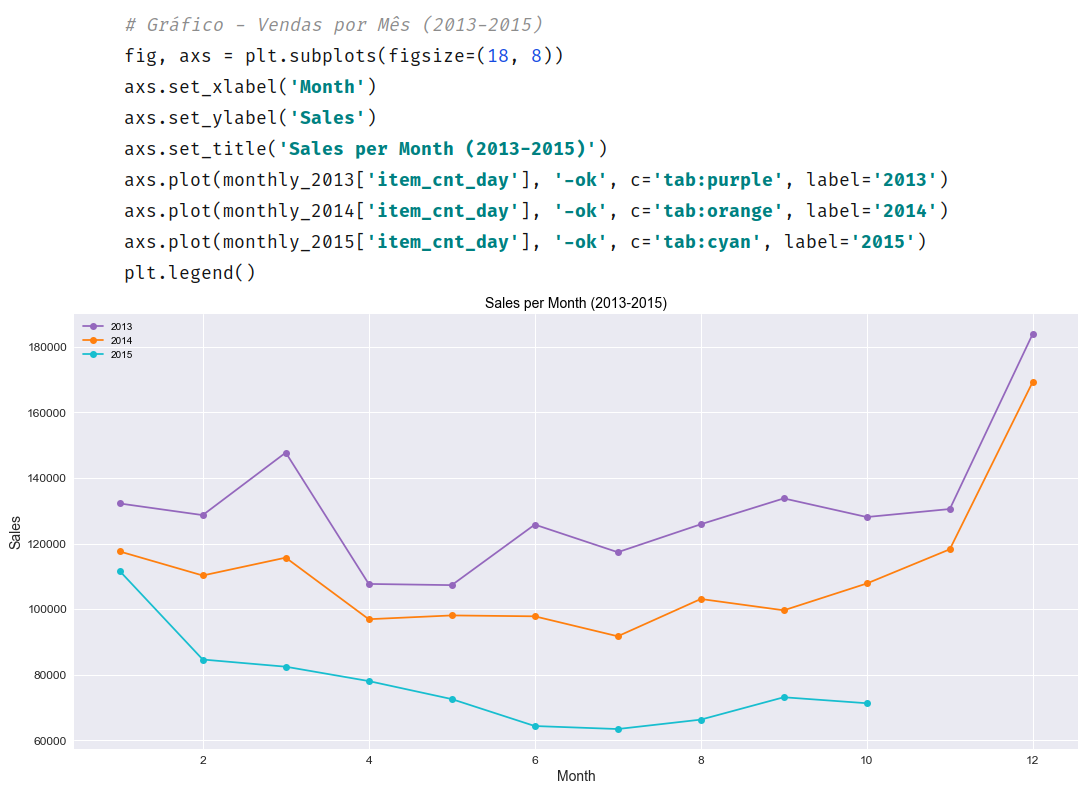
# 

# 4. Análise e Exploração dos Dados

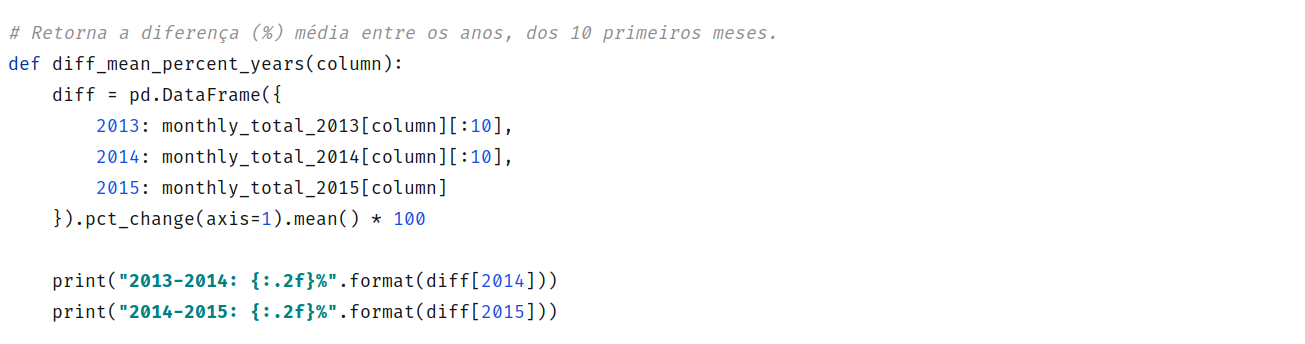
Após a realização do Processamento/Tratamento dos Dados, análises foram elaboradas, com objetivo de obter *insights* e identificar padrões no *dataset* “*sales*”.



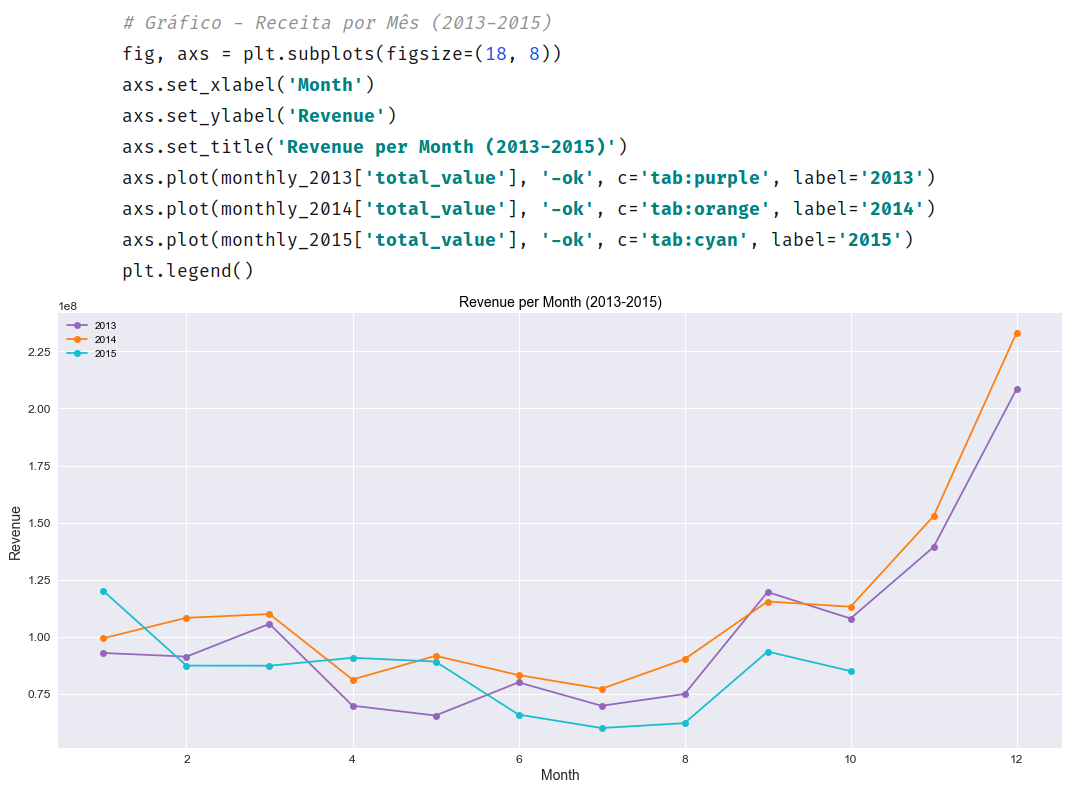
Para obter uma visão do desempenho das **vendas** e da **receita** nos anos de 2013, 2014 e 2015, foram construídos gráficos de linhas, a fim de permitir a comparação entre os anos. Portanto, fez-se necessário, o somatório mensal das vendas e da receita, de cada ano.



Conforme apresentado no Gráfico “*Sales per Month (2013-2015)*”, com o decorrer dos anos, as vendas diminuíram significativamente. Para obter o percentual referente à variação das vendas e da receita, foi calculado a diferença média entre os anos, dos 10 primeiros meses (Janeiro a Outubro). No cálculo realizado, não foi levado em consideração os meses de Novembro e Dezembro, pois no ano de 2015, os dados referente a tais meses não estavam presentes no *dataset*.



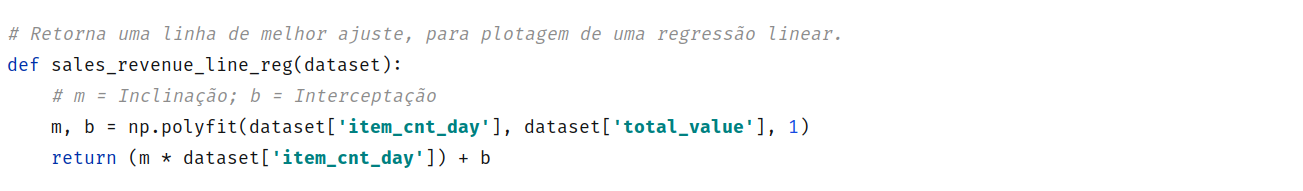
Após a execução do cálculo, identificou-se que houve uma queda média de 16,92% nas vendas entre os anos de 2013 e 2014. Entre os anos 2014 e 2015, a queda nas vendas foi maior que o período anterior, alcançando em média 26,37%.



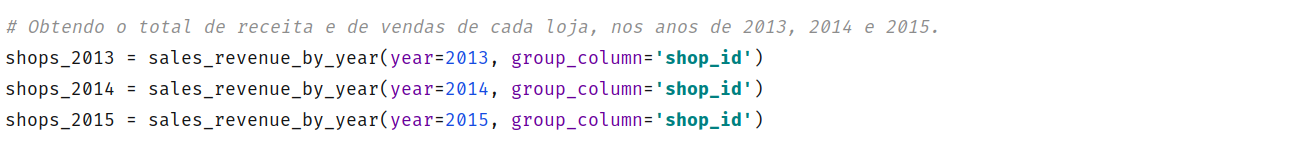
Com o Gráfico “*Revenue per Month (2013-2015)*”, percebe-se que a receita permaneceu quase estável durante os meses de Janeiro e Outubro. Após realizar o cálculo da diferença média, verificou-se que, houve um aumento médio de 12,24% entre os anos de 2013 e 2014. Entretanto, entre os anos de 2014 e 2015, um decréscimo médio de 12,82%.

Após a análise das vendas e da receita da empresa como um todo, foram elaborados gráficos de dispersão para verificar a distribuição anual das vendas e da receita, relacionada às lojas e às categorias de produtos.

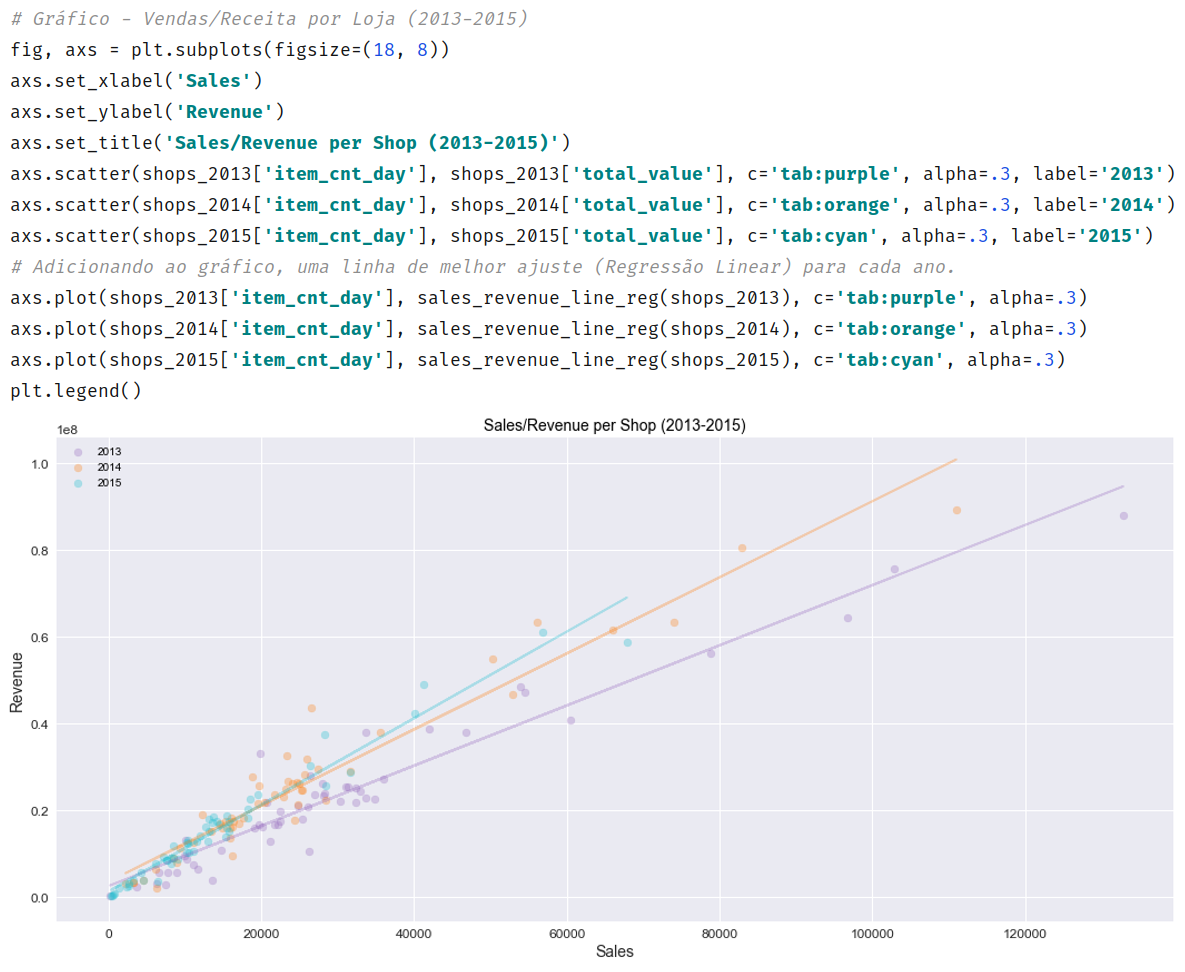
Além disso, foi incluído aos gráficos, para cada ano, uma linha para representação da distribuição dos dados, por meio de Regressão Linear, a fim de facilitar a interpretação dos resultados.



Após a definição da função responsável por retornar a Regressão Linear, obteve-se o total de vendas e receita, relacionada à cada loja, de cada ano.

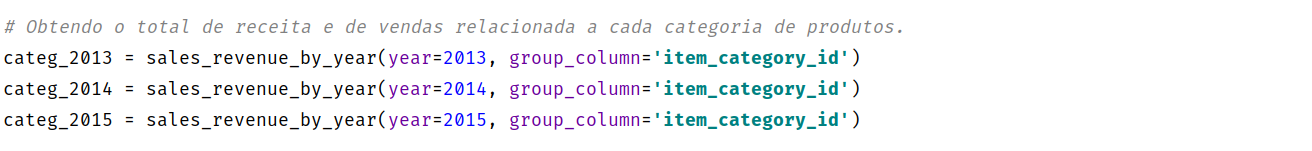


Por meio do Gráfico “*Sales/Revenue per Shop (2013-2015)*”, é possível observar que, com o passar dos anos, existe uma tendência de queda nas vendas das lojas. Entretanto, há a tendência contrária (de aumento) com relação à receita gerada pelas lojas.

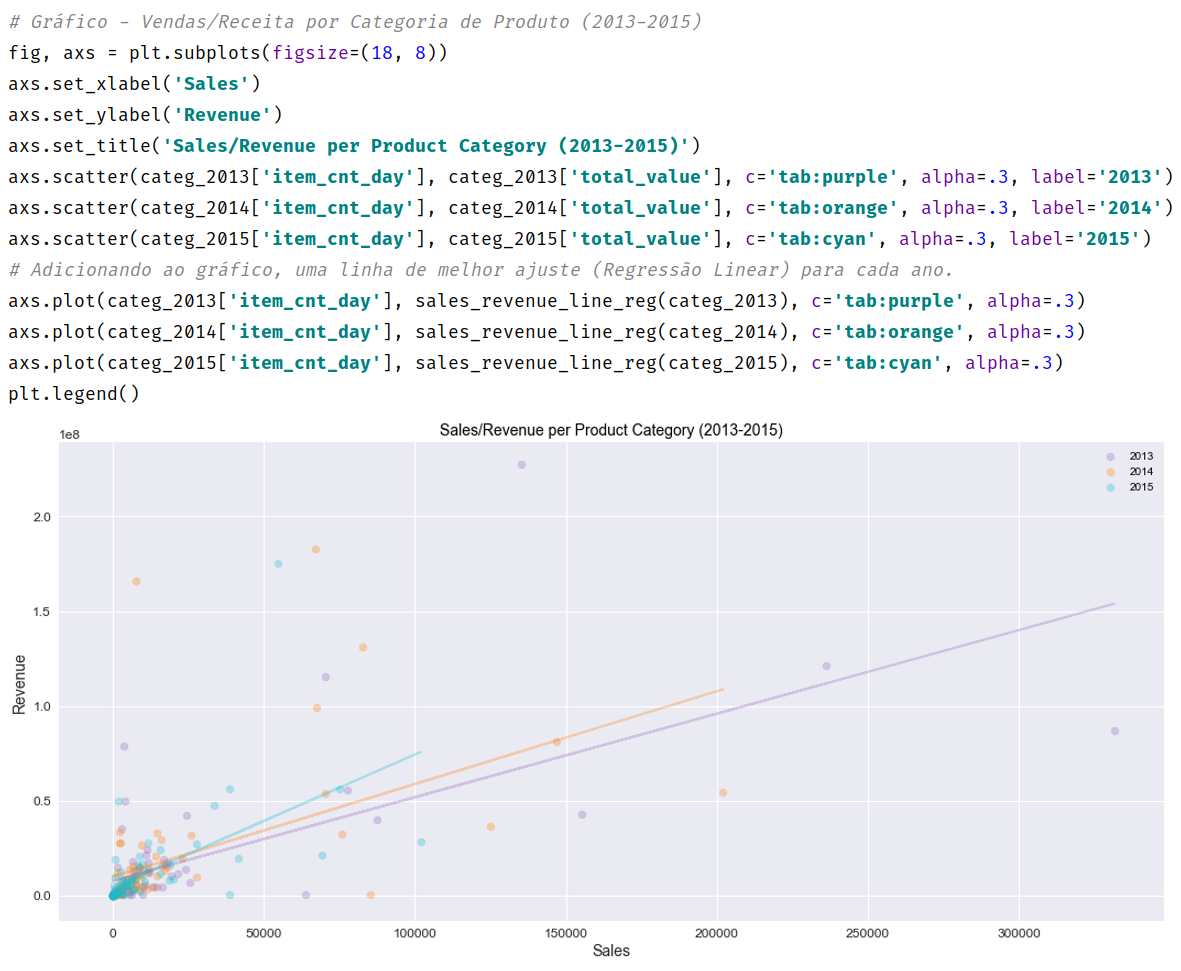


Além disso, observa-se que durante os anos, existem dois grupos, que possuem características similares e representam a maioria das lojas da empresa. O primeiro grupo, realizou até 20.000 vendas, com geração de receita de até 20.000.000. Enquanto o segundo grupo, realizou entre 20.000 e 40.000 vendas e com geração de receita entre 20.000.000 e 40.000.000.

Após as análises relacionadas ao desempenho anual das lojas, obteve-se o total anual de vendas e receita de cada categoria de produtos, assim como, a representação da distribuição dos dados, por meio de Regressão Linear.



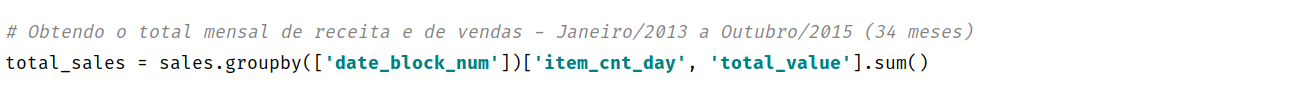
Por meio do Gráfico “*Sales/Revenue per Product Category (2013-2015)*”, é possível observar que, com o passar dos anos, existe uma tendência de crescimento na receita, quando associada às categorias de produtos. Entretanto, há a tendência contrária (de queda) com relação às vendas, quando associadas às categorias de produtos.



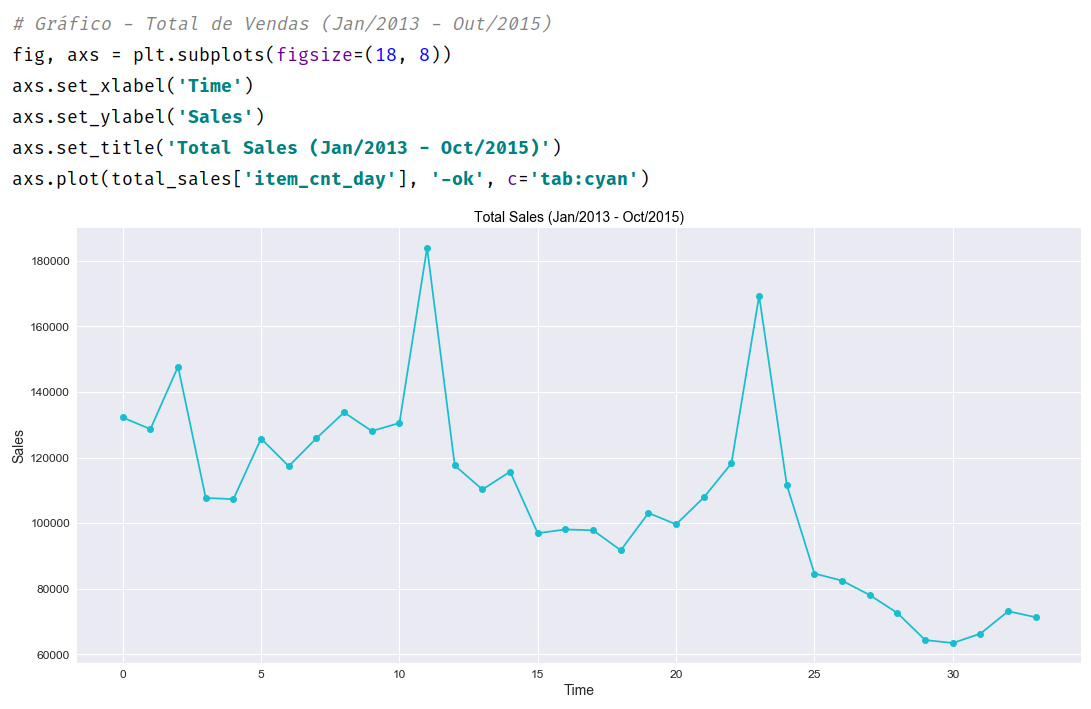
Além disso, observa-se que, nos anos de 2013 e 2014, houveram categorias de produtos que representaram aumentos significativos nas vendas e na geração de receita. Por exemplo, em 2013, algumas dessas categorias estiveram relacionadas a mais de 300.000 das vendas realizadas, e outras, relacionadas à geração de receita de mais de 200.000.000.

Devido ao fato de que, os registros de vendas de 2015 estavam limitados aos meses de Janeiro a Outubro, é provável que as tendências identificadas nos gráficos de dispersão foram influenciadas pela ausência de registros de vendas referentes aos meses de Novembro e Dezembro, principalmente quando se comparado aos anos anteriores.

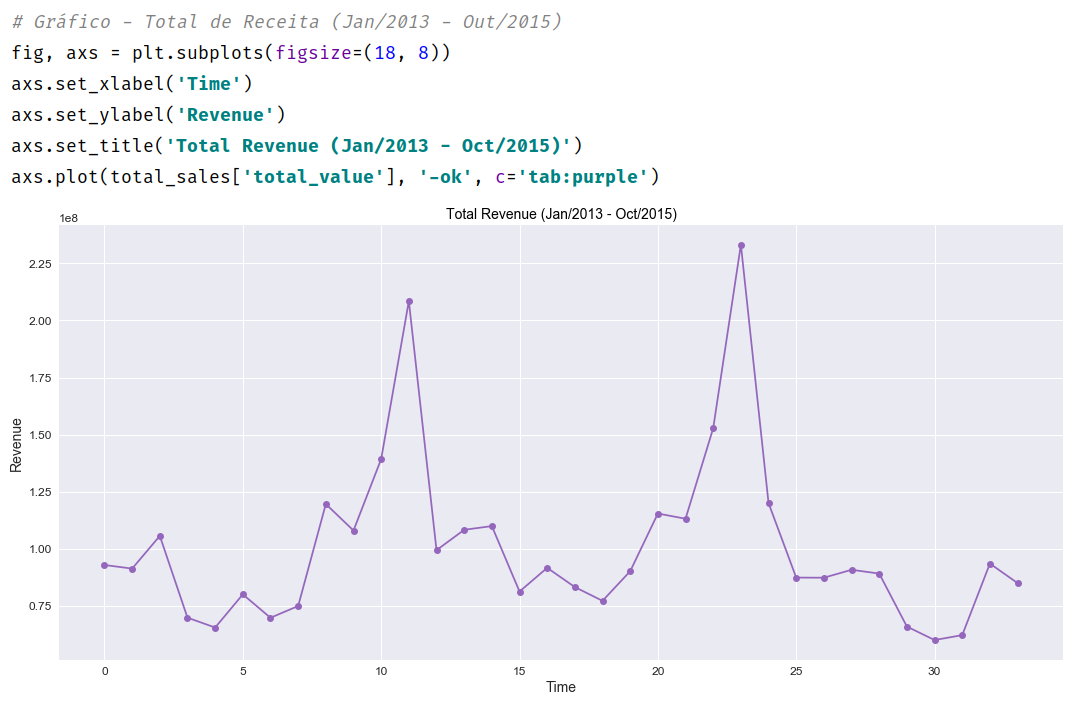
Após a análise dos dados e dos gráficos relacionados às vendas e à receita de cada ano, foi examinado o desempenho geral da vendas e da receita, de modo a observar as **tendências** e as **sazonalidades**. Para tal, foi necessário obter o somatório total mensal das vendas e da receita.



Com o resultado do agrupamento, tornou-se possível a construção de dois gráficos de linhas. O primeiro, relacionado às vendas, enquanto o segundo, relacionado à receita gerada. Ambos os gráficos representam o desempenho da empresa no decorrer dos três anos, entre Janeiro de 2013 a Outubro de 2015, totalizando 34 meses.

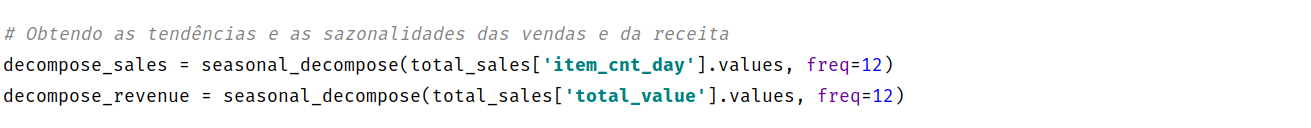


Conforme apresentado no Gráfico “*Total Sales (Jan/2013 - Oct/2015)*”, é possível identificar a queda nas vendas com o passar do tempo, assim como indicado nos gráficos anteriores. Vale ressaltar que, existem picos elevados nas vendas, ocorridos nos meses de Dezembro de 2013 e 2014.



No Gráfico “*Total Revenue (Jan/2013 - Oct/2015)*”, a receita demonstra estar mais estável, se comparado com as vendas, assim como mencionado anteriormente, visto que as diferenças médias entre os anos de 2013, 2014 e 2015, praticamente se anulam.

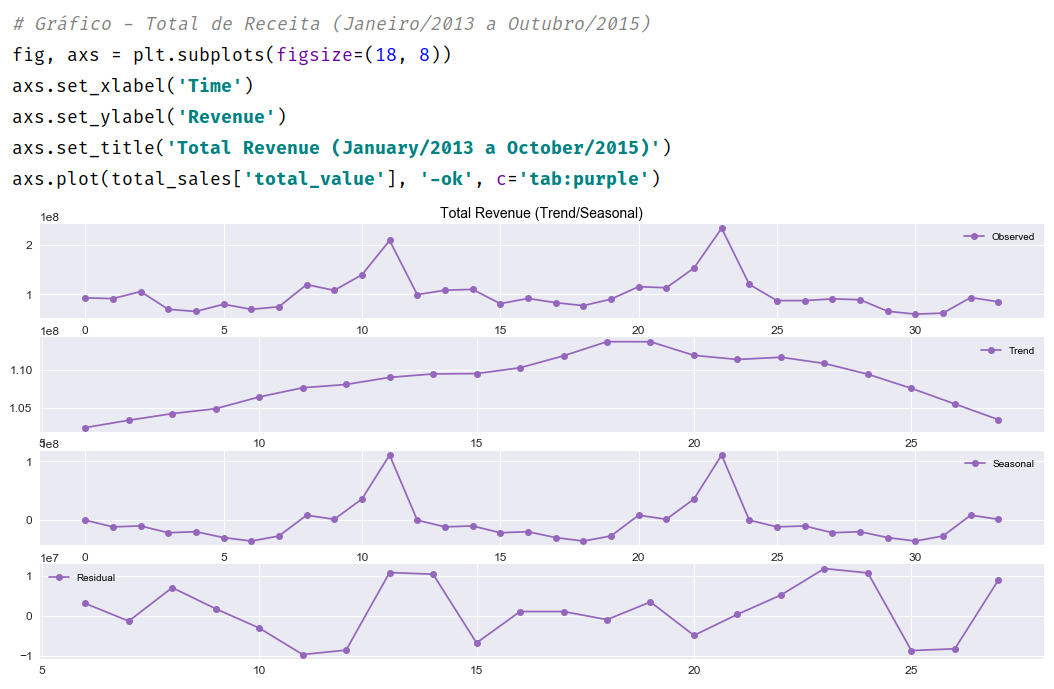
Em seguida, a partir da função “*seasonal\_decompose*”, disponível através da biblioteca “*statsmodel*”, obteve-se informações sobre a tendência, a sazonalidade e os resíduos, referentes aos dados do total de vendas e receitas.



No Gráfico “*Total Sales (Trend/Seasonal)*”, é possível observar que, tornou-se mais evidente, a existência de uma tendência de queda nas vendas da empresa. Além disso, percebe-se que há uma sazonalidade anual, ou seja, que os fenômenos (vendas) se repetem, de forma idêntica, a cada 12 meses.



Apesar de que, no Gráfico “*Total Sales (Trend/Seasonal)*” é perceptível a existência de uma sazonalidade (anual), há uma maior dificuldade quanto à interpretação da tendência, referente à geração de receita da empresa.

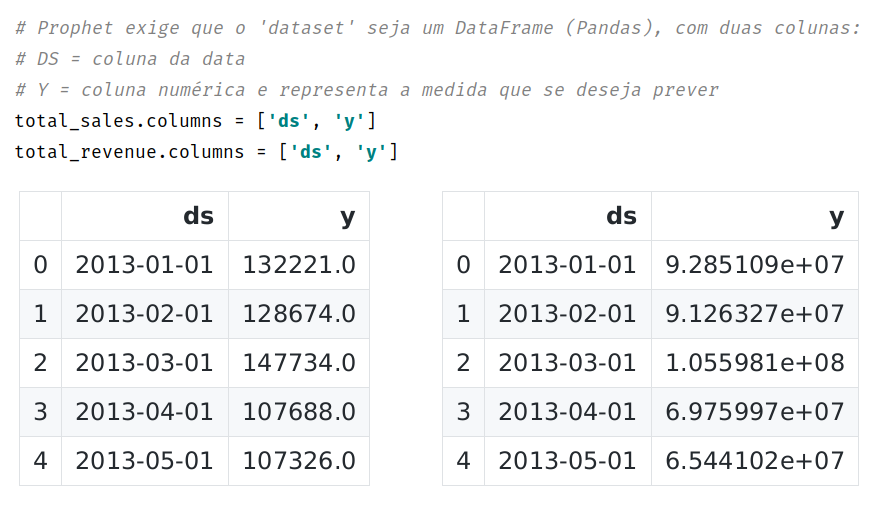


# 5. Modelo de *Machine Learning*

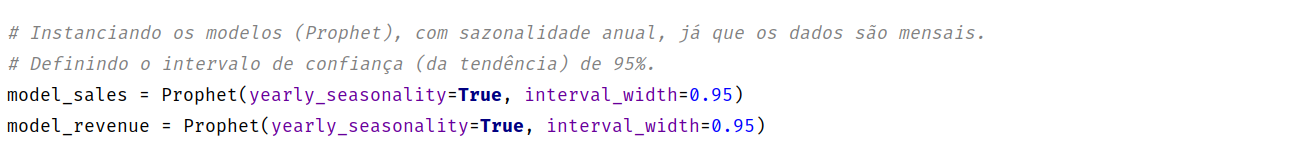
Para a criação do modelo de *Machine Learning* e realização da previsão de vendas, foi utilizado a biblioteca *Prophet*[[2]](#footnote-1), criada pela equipe *Core Data Science*[[3]](#footnote-2) (CDS), do *Facebook*.

*Prophet* é um procedimento para previsão de séries temporais com base em um modelo aditivo, no qual tendências não lineares se ajustam à sazonalidade anual, semanal e diária, além de efeitos de férias. Além disso, *Prophet* é uma biblioteca de código aberto, implementado nas linguagens R e Python.

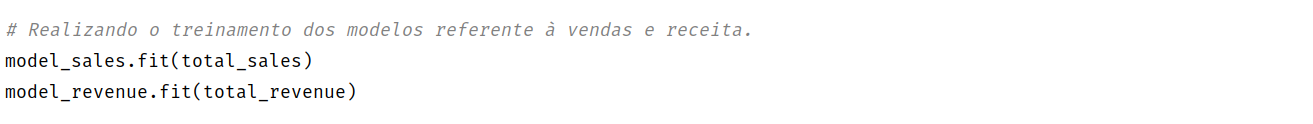
Após o pré-processamento dos dados necessários para criação do modelo (com *Prophet*), apresentado no capítulo [3.2](#_3xt2g1wb77xd), era preciso que o nome das colunas dos datasets “*total\_sales*” e “*total\_revenue*” fossem renomeadas, pois o *Prophet* exige que o *DataFrame* tenha obrigatoriamente duas colunas, “*DS*”, que representa a coluna da data, e “*Y*”, representando a coluna numérica, a qual deseja-se prever.



Para criação dos modelos de previsão com o *Prophet*, considerou-se que a sazonalidade é anual, já que os dados são mensais, e para a tendência, foi definido um intervalo de confiança de 95%, pois o intervalo padrão do *Prophet* é 80%.

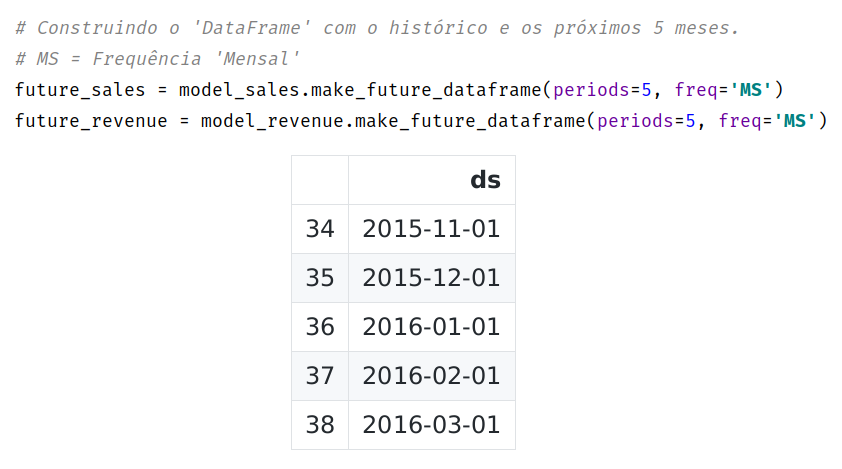


Uma vez que os *datasets* atende ao formato exigido e os modelos já foram instanciados, é realizado a etapa de treinamento dos modelos com os *datasets* informados.



Após o treinamento, o *Prophet* permite construir um novo *DataFrame*, que inclui o histórico e as datas dos períodos futuros. Neste caso, a frequência informada é “Mensal”, podendo assim, incluir os próximos 5 meses, a quais se referem à Novembro de 2015 a Março de 2016.

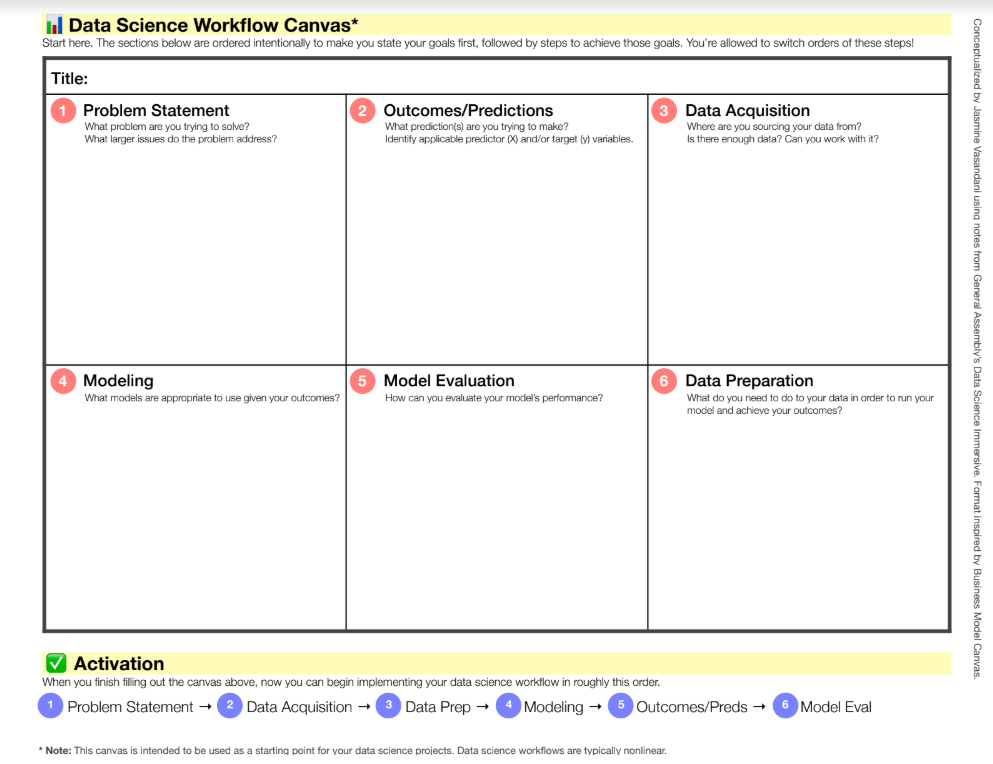
# 



[...]

# 6. Apresentação dos Resultados

Nessa seção você deve apresentar os resultados obtidos. Apresente gráficos, dashboards, conte a sua história de forma bastante criativa. Aqui você pode utilizar os modelos de Canvas propostos por Dourard (clique [aqui](https://www.louisdorard.com/machine-learning-canvas)) ou por Vasandani (clique [aqui](https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0)).



# 7. Links

[...]

# REFERÊNCIAS

Um projeto de Ciência de Dados não requer revisão bibliográfica. Portanto, a inclusão das referências não é obrigatória. No entanto, caso você deseje incluir referências relacionadas às tecnologias ou às metodologias usadas em seu trabalho, relacione-as de acordo com o modelo a seguir.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

1. *Kaggle*, uma subsidiária da *Google LLC*, é uma comunidade on-line de cientistas de dados e profissionais de aprendizado de máquina. [↑](#footnote-ref-0)
2. <https://facebook.github.io/prophet/> [↑](#footnote-ref-1)
3. <https://research.fb.com/teams/core-data-science/> [↑](#footnote-ref-2)