**Afinal, como se desenvolve um projeto de Data Science?**

O termo Data Science se tornou um *hot topic* mundial na indústria da tecnologia. O rápido avanço computacional tem permitido análises de quantidades de dados cada vez maiores, possibilitando a descoberta de padrões e *insights*até em tempo real.

Com isso, uma pergunta comum que ouvimos de amigos ou pessoas interessadas na área é: *“Cara, Data Science parece muito legal! Mas por onde começar?”.*Nesse post, daremos uma visão geral de seis etapas que fazem parte do ciclo de vida de um projeto na área: **entendimento do problema, coleta de dados**,**exploração dos dados**,**análise profunda dos dados**,**comunicação dos resultados**e**feedback.**

Você já deve ter começado a notar que essas etapas são similares a projetos de outras áreas, inclusive fora da TI, certo? O ciclo de vida de um projeto de DS envolve fases semelhantes ao ciclo de resolução de problemas do dia a dia de qualquer profissional. Por exemplo, vamos supor uma situação na qual o objetivo é comprar uma televisão. Por onde você começaria?

1. Preciso de uma televisão > ***Surgimento da demanda***
2. Eu realmente preciso de uma televisão nova? Quanto de vou precisar gastar? >***Entendendo o problema***
3. Preciso pesquisar quais os modelos e tecnologias envolvidas das televisões da atualidade *>****Coleta de dados***
4. Parece que alguns modelos ainda não estão sendo vendidos no meu país. Meu videogame tem suporte a esse modelo? Esse modelo é considerado uma Smart TV? > ***Processamento e exploração de dados/Análise de dados***
5. Encontrei um modelo de televisão que atende minha necessidade. > ***Comunicação de resultados***
6. Instalei a televisão e estou pronto para testá-la. > ***Feedback***

Faz sentido, né?! Vamos então entender cada uma dessas etapas. Mas, antes de começarmos, vale a pena ressaltar que o ciclo de vida de um projeto de Data Science pode conter algumas diferenças de empresa para empresa dependendo de vários fatores, restrições e recursos disponíveis. Algumas etapas podem, inclusive, serem removidas ou adicionadas de acordo com o negócio.



**1. Entendendo o problema**

Essa é uma das etapas que considero uma das mais importantes de todo o ciclo. É nela que precisamos gastar tempo suficiente para entender o problema de forma mais clara possível. Para isso, é importante que estejamos em constante comunicação com os *stakeholders*, as pessoas envolvidas no projeto, e/ou aqueles que irão se beneficiar com a solução.

Nessa fase, é papel do Cientista de Dados entender as dores dos *stakeholders*e fazer as perguntas certas, antes mesmo de “colocar a mão na massa”.

***Dica 1****:* Utilizar a[Técnica dos 5-Ws](https://its.unl.edu/bestpractices/remember-5-ws):

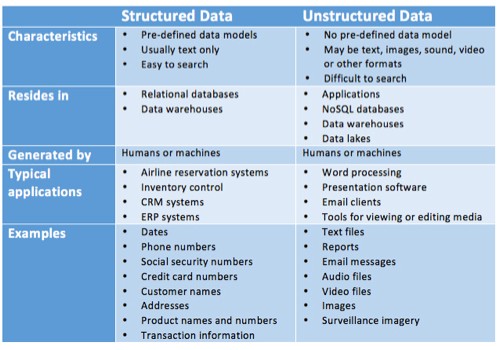
* **Porquê?** (*Why?*): Porque é importante essa análise para o negócio?
* **Quem?**(*Who?*): Quem iremos analisar? Nossos compradores? Fornecedores?
* **O quê?** (*What?*): O que iremos analisar? Comportamento de compra?
* **Onde?** (*Where?*): A análise estará voltada para o contexto nacional ou internacional?
* **Quando?** (*When?*): Qual período será considerado para as análises?

***Dica 2:*** Utilizar a [Técnica dos 5 Porquês](https://www.isixsigma.com/tools-templates/cause-effect/determine-root-cause-5-whys/).

**2. Coleta de dados**

Uma vez definido o problema, precisamos começar a extrair e coletar os dados. Nessa etapa, é fundamental entender quais os tipos de dados irão pautar nosso projeto:

* ***Dados internos*** (presentes em bancos de dados, planilhas, etc.) **x** ***Dados Externos*** (bases de dados públicas ou pagas, etc.)
* ***Dados estruturados*** (tabelas dos nossos DBs) ***x Dados não-estruturados*** (conteúdos de redes sociais, de sites externos, etc.).



*(Fonte: [Datamation](https://www.datamation.com/big-data/structured-vs-unstructured-data.html" \t "_blank))*

Esse mapeamento irá auxiliar na decisão das tecnologias que utilizaremos para coletar nossos dados (consultas SQL, *crawlers*, APIs, etc.).

**3. Processamento/Tratamento de dados**

Já coletamos os dados mas precisamos tratá-los antes de começar nossas análises. Nessa etapa é necessário estar atento a registros duplicados, faltantes, formatados de forma não-convencional (ex.: campos de data), inválidos (ex.: idade negativa), inconsistências de cadastros (ex.: data da venda anterior a data de início de venda de um produto).

Após essa identificação, é importante pensar na melhor forma de contornar de acordo com as regras do negócio. Por exemplo, no caso de um registro onde o valor da compra está vazio, você poderia preenchê-lo com zero, ou com a média dos valores de compra de produtos similares, etc.

***Dica 1:*** A biblioteca Python [missingno](https://github.com/ResidentMario/missingno" \t "_blank) te ajuda a encontrar valores faltantes/nulos nos seus dados.

***Dica 2:*** A biblioteca Python [pandas-profiling](http://l/)te ajuda a explorar seus dataframes com informações estatísticas, descritivas, histogramas e etc.



Biblioteca *missingo*(Fonte: [GitHub/ResidentMario](https://github.com/ResidentMario/missingno))

**4. Exploração de dados**

Agora sim, a etapa do ciclo de vida do nosso projeto de Data Science onde a resolução do problema inicial começa a tomar forma!

Na etapa de exploração de dados precisaremos lembrar do *Passo 1: Entendimento do Problema*. Nessa fase, se intensifica a necessidade de habilidades analíticas e criativas para pensar em ideias e hipóteses a serem validadas. É importante que você busque identificar padrões interessantes nos seus dados (lembra do estudo da [estudo famoso da fralda e da cerveja](http://computerworld.com.br/blog/opiniao/2013/12/04/fraldas-cervejas-e-big-data)?!).

Existem várias ferramentas e bibliotecas em várias linguagens de programação que podem auxiliar nessa etapa. Dá uma conferida abaixo:

Ferramentas open-source:

* [Jupyter Notebook](http://jupyter.org/)
* [Metabase](https://www.metabase.com/)
* [Weka](https://www.cs.waikato.ac.nz/ml/weka/)
* [R Shiny](https://shiny.rstudio.com/)

Ferramentas gratuitas:

* [Power BI Desktop](https://powerbi.microsoft.com/pt-br/desktop/)
* [Qlik Sense Desktop](https://www.qlik.com/pt-br/products/qlik-sense/desktop)
* [Tableau Desktop](https://www.tableau.com/pt-br/products/desktop)

Bibliotecas:

* [Pandas](https://pandas.pydata.org/)
* [NLTK](https://www.nltk.org/)

**5. Análise profunda de dados**

Do inglês *Perform in-depth analysis,*essa etapa pode não estar presente em todos os projetos de Data Science. É nessa fase que realizamos passos de seleção de *features*, implementamos e aplicamos modelos estatísticos e de *Machine Learning*para validar hipóteses.

Suponha que seu problema seja criar um modelo preditivo que identifique quando um cliente deixará sua plataforma. Precisaremos então criar um modelo mais complexo e automatizados para que os *stakeholders* tenham informações suficientes para tomar decisões sobre possíveis *churns*.

*“- Mas porque é uma etapa as vezes ausente?”* Porque alguns problemas são resolvidos na etapa anterior, de exploração de dados, e não exigem a aplicação de uma metodologia mais complexa para sua resolução.

Para nos auxiliar nessa etapa, contamos com bibliotecas como a [scikit-learn](http://scikit-learn.org/stable/" \t "_blank), que encapsula vários modelos de Classificação, Regressão, Clusterização, Redução de dimensionalidade e etc. Além disso, algumas ferramentas estão começando a disponibilizar plugins para análises mais complexas, como a opção de [forecasting do Power BI](https://powerbi.microsoft.com/pt-br/blog/tag/forecasting/" \t "_blank).

**6. Comunicação de resultados e Feedback**

O ciclo parece completo, certo? Ainda não…

O sucesso de um projeto de Data Science (como qualquer outro) depende da comunicação efetiva dos resultados, que darão suporte ao processo de tomada de decisão no negócio. Para isso, use e abuse do *storytelling*, que nada mais é do que a capacidade de contar boas histórias para seus *stakeholders*, mostrando como os resultados podem ajudá-los decidir na tomada de decisões.

Além disso, é importante também que o Cientista de Dados se preocupe com a atualização dos resultados, de acordo com a volatilidade do negócio. Por isso é crucial que todo o desenvolvimento do projeto seja elaborado de forma reproduzível.

**Conclusão**

De forma geral, esse é o ciclo de vida de projetos de Data Science. Algumas etapas podem ser adicionadas ou removidas dependendo da necessidade de cada negócio.

Você conseguiu pensar em alguma etapa importante do ciclo de vida que não comentamos por aqui, ou outras ferramentas/bibliotecas que podem agregar a alguma etapa? Deixe seu comentário aqui embaixo!