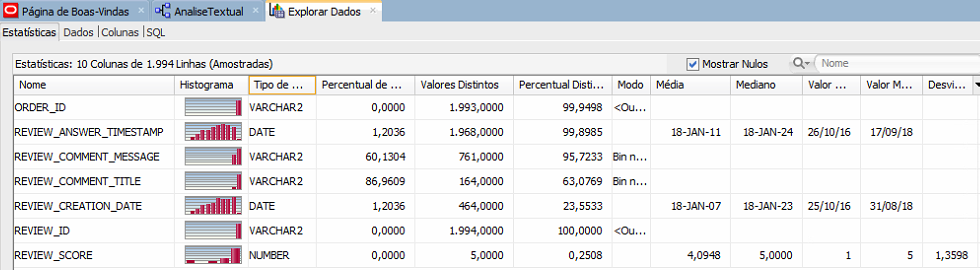
**Cientista de dados**

* Que tal uma análise textual dos clientes que deixaram comentários sobre suas compras?

1. Higienização da base “olist\_order\_reviews\_dataset”.
2. Carga da base em banco oracle (será utilizado o Oracle Data Miner)
3. Exploração dabase (verificar como os dados estão distribuídos)



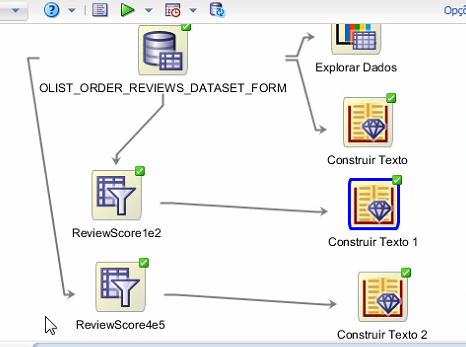
Algumas observações:

-Aproximadamente 2 milhões de reviews

-Aproximadamente 60% dos comentários são nulos e 96% são distintos

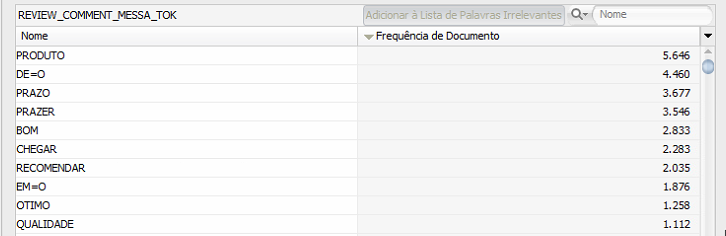
-Média do Score de Review aproximadamente 4,1

4- Aplicando o Text Mining

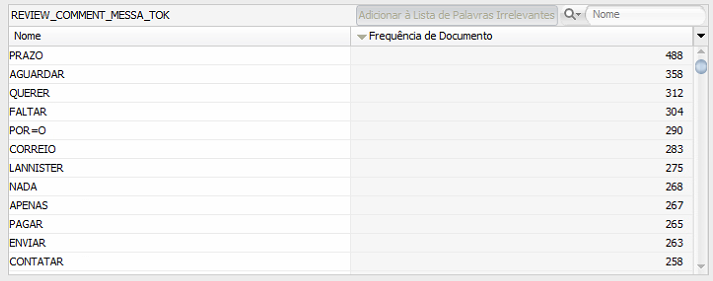


Resultados:

-Para reviews com avaliação boa (4 e 5) se destacam as palavras Prazo, Bom, Ótimo, Qualidade



-Para reviews com avaliação baixa (1 e 2) se destacam as palavras Prazo, Aguardar, Correio, Lannister (loja).



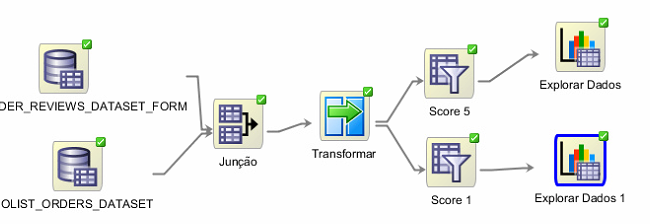
**Em ambos os casos (avaliação positiva e negativa) a principal palavra foi com relação ao cumprimento do prazo de entrega.**

* Alguns clientes não escreveram um comentário. Mas por que eles estão satisfeitos?

1- Junção da tabela de “reviews” com a “orders”.

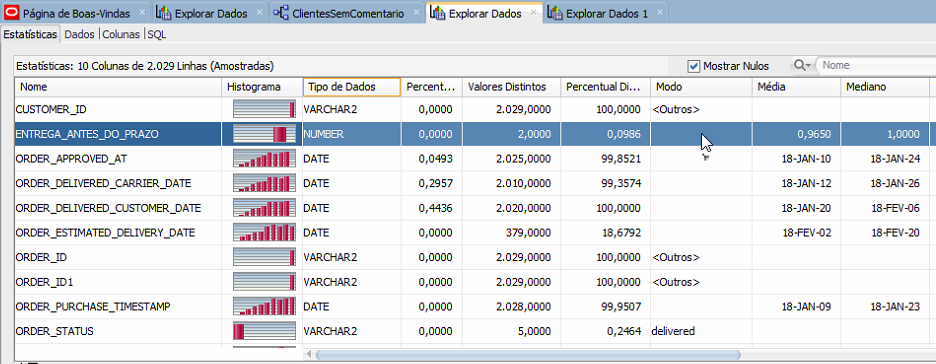
2- Criei um campo (Entrega antes do prazo) comparando a data prevista de entrega com a data real de entrega.

3- Fiz dois filtros para comparar os resultados dos que não comentaram e deram score 5 e dos que não comentaram e deram score 1.



4-Exploração Score 5 (sem comentários):

Mais de 96% das entregas foram antes ou dentro do prazo.



5- Exploração Score 1 (sem comentários):

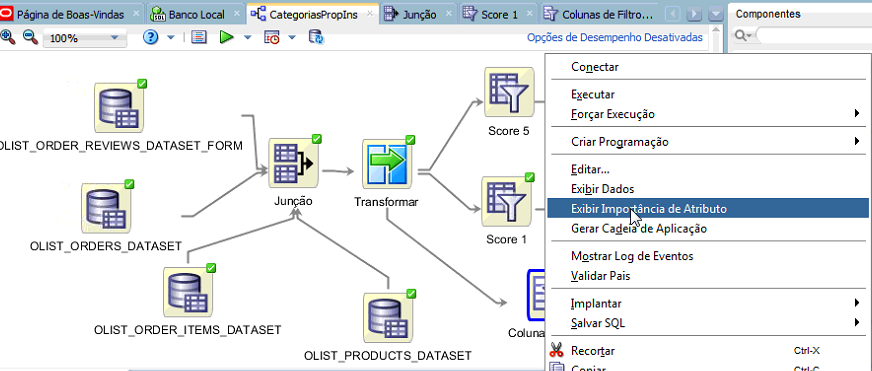
Menos de40% das entregas foram dentro do prazo.

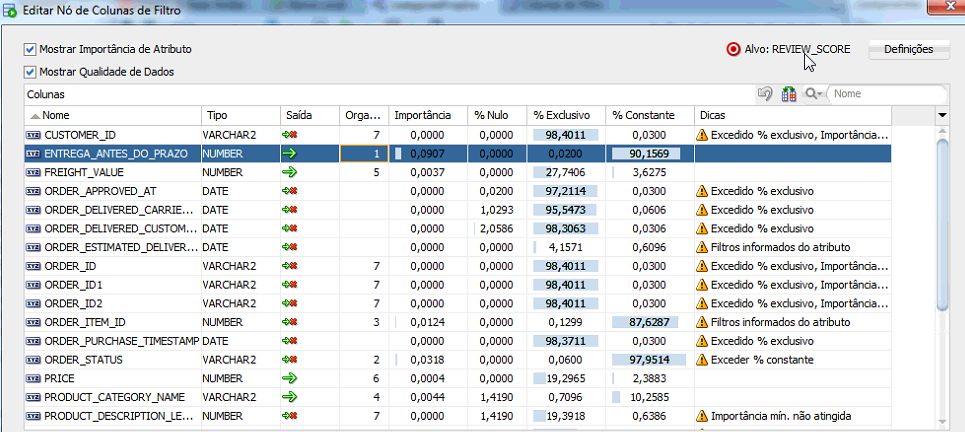


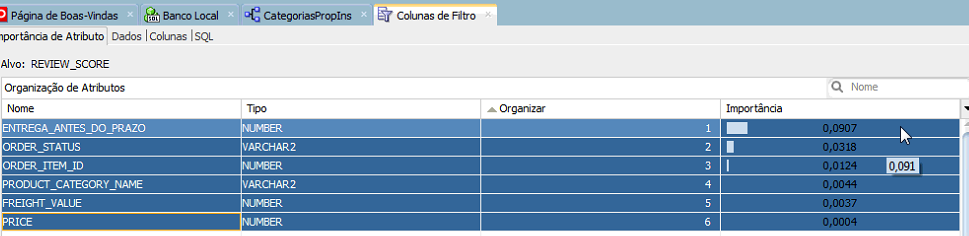
**A grande maioria (96%) dos clientes satisfeitos (score 5) e que não escreverem comentários tiveram suas entregas realizadas antes ou dentro do prazo.**

* Divirta-se descobrindo as categorias de produtos mais propensas à insatisfação do cliente.

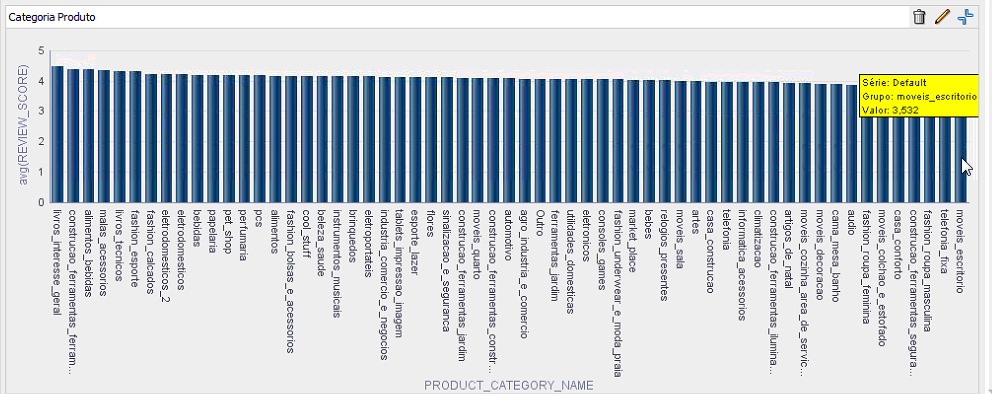
Foi montado um outro workflow no dataminer e executei uma analise para verificar a importancia dos atributos levando como alvo o review score. Foi constatado que a entrega no prazo teve o maior peso para as variações do score review. Categoria do produto aparece na lista também.



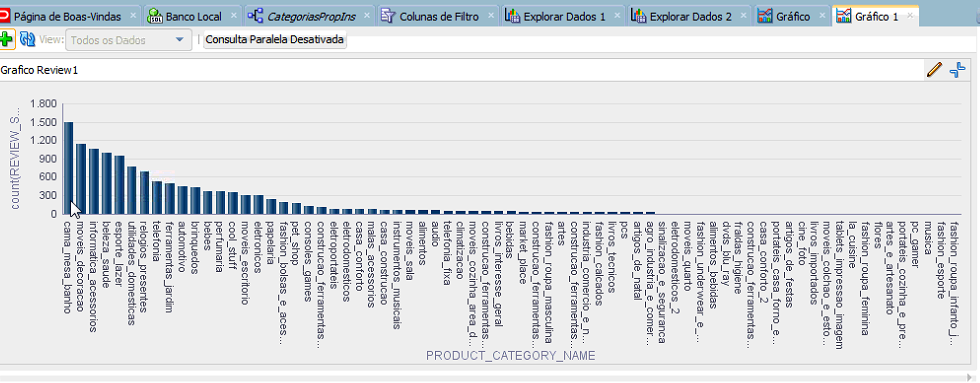




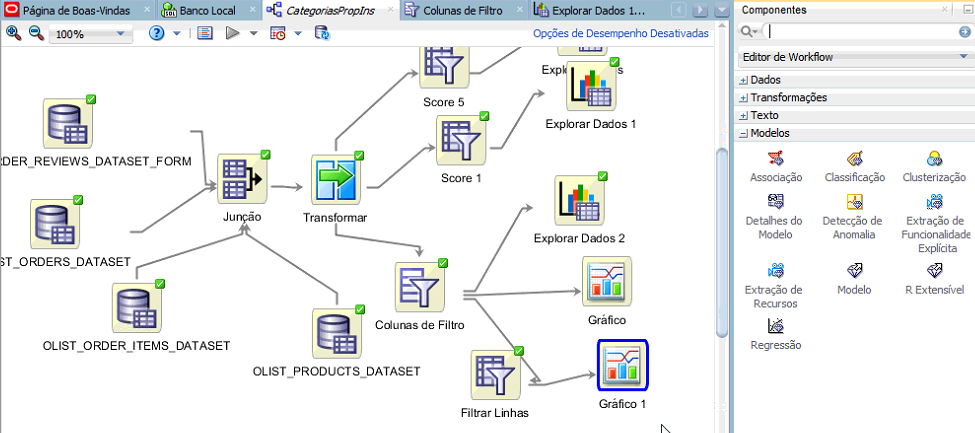
Com a média dos reviews, a categoria que teve a menor média foi a "móveis de escritório"



Filtrando os reviews que obtiveram "score 1", a categoria que obteve mais reviews com essa nota foi "cama\_mesa\_banho".



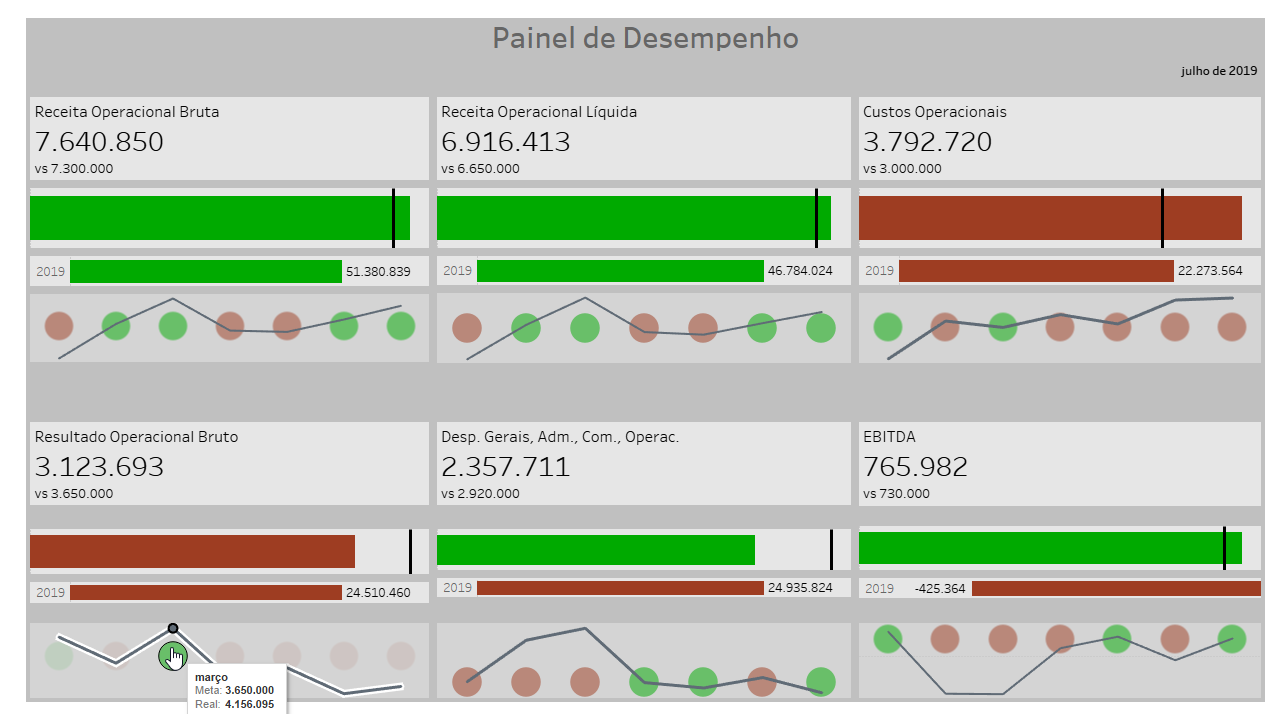
Esse é uma visualização da ferramenta e os modelos de algoritmos que possui (Associação, Classificação, Clusterização, Detecção de Anomalia, etc...)



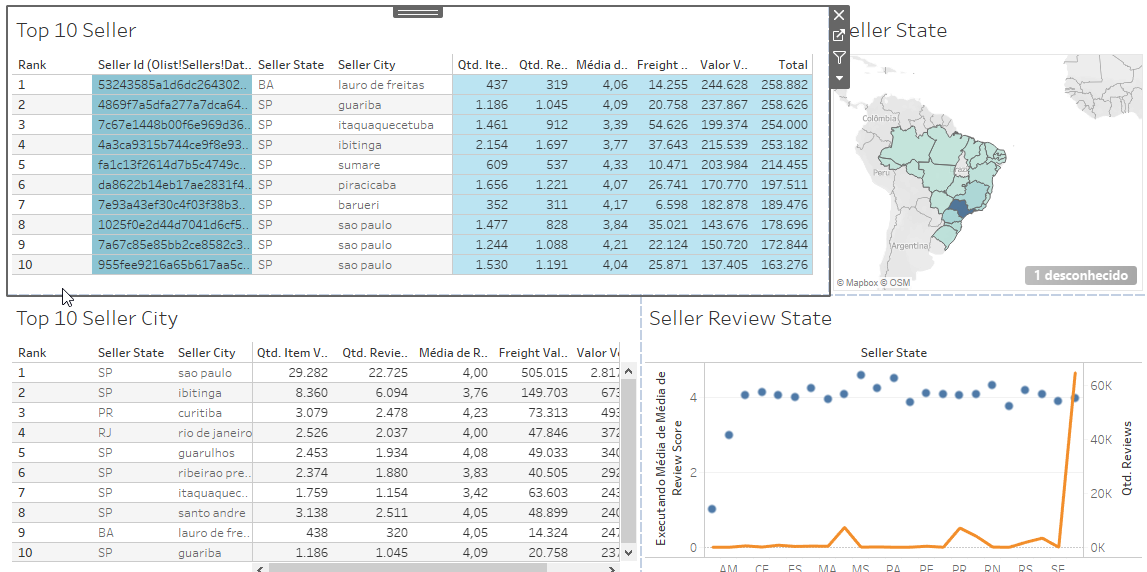
**Analista de dados e Business Intelligence**

* Pense em alguns KPIs para monitoramento. Talvez outros para direcionamento dos gestores!

Dashboard de monitoramento eu pensei em fazer algo parecido com o modelo abaixo, mas senti falta de informações de meta ou de um objetivo de referência. Esse modelo abaixo é de um outro projeto que realizei.

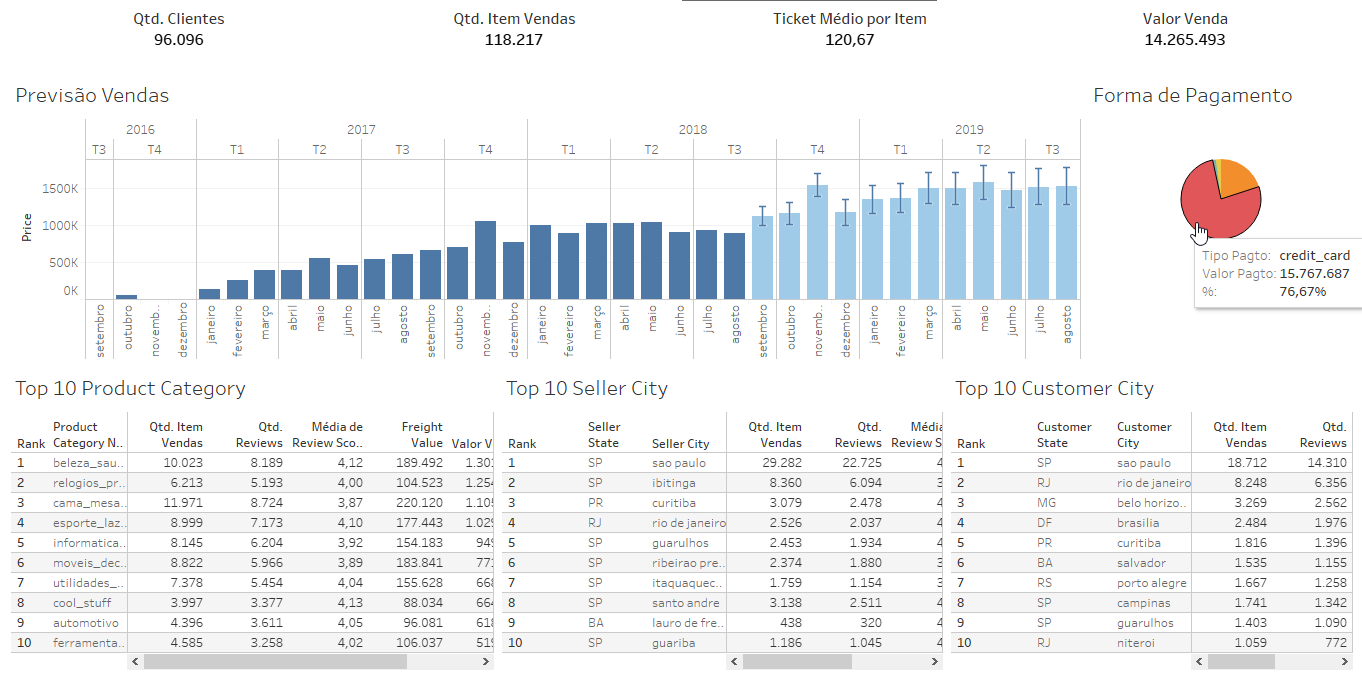


* Um cruzamento dos dados poderia gerar relatórios interessantes. Afinal, quem são os Top 10 em vendas? Que tipo produtos eles vendem? Qual é o impacto deles para o negócio?



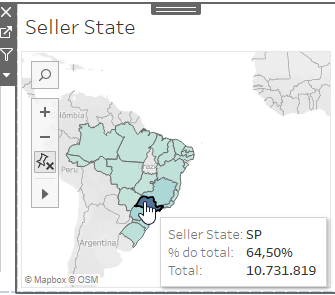


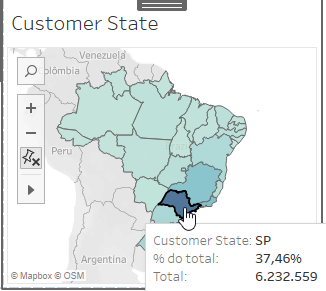
* Que tal realizar uma análise exploratória dos dados. E então? Algo lhe chama a atenção?



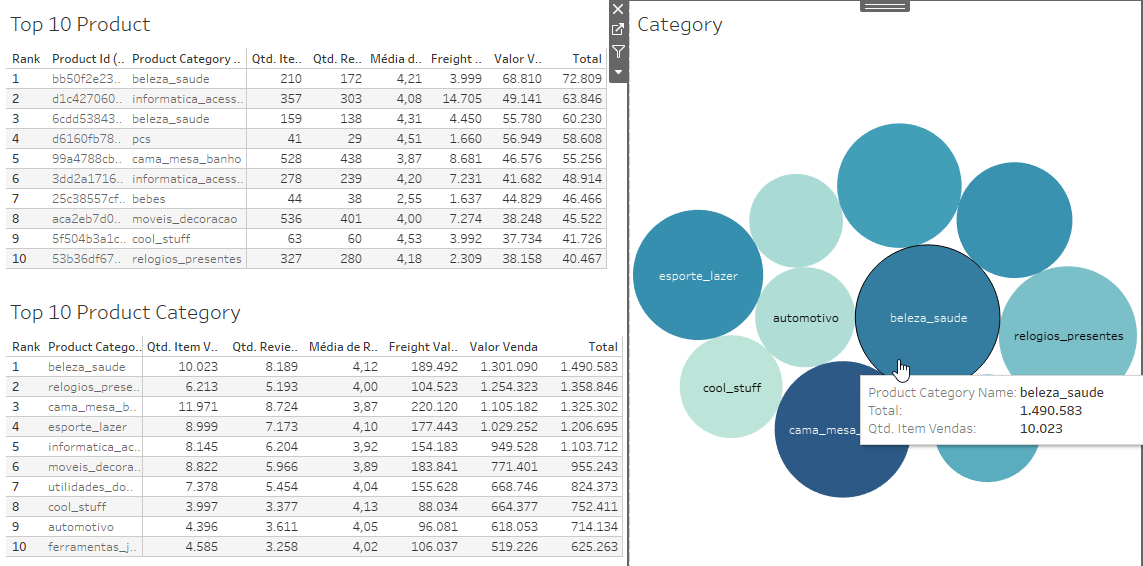
-Mais de 96 mil clientes, 118 mil itens vendidos, ticket médio de 120,67, valor de venda maior que 14M.

-Mais de 76% dos pagamentos são por cartão de crédito (20 % Boleto).

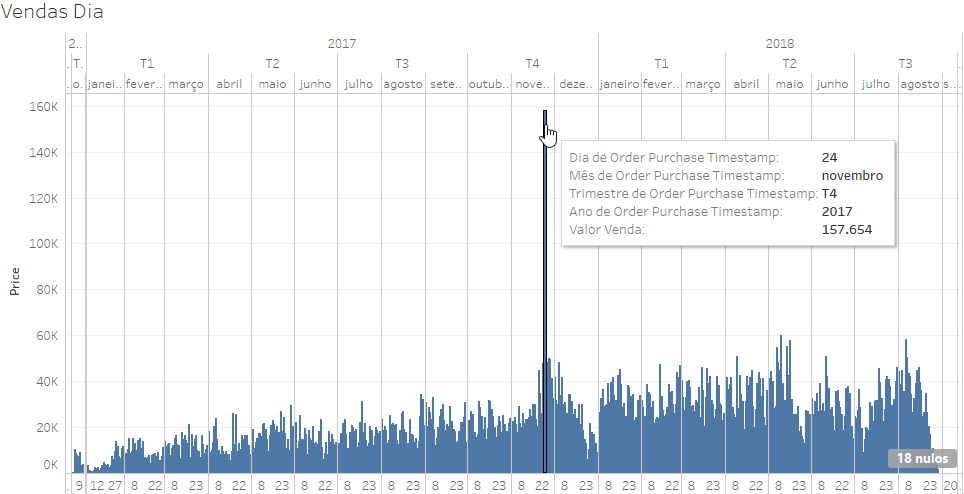




- Vendedores do estado de SP representam 64,5 % de toda a venda. Os compradores já representam 37,46% do total.



A Categoria de Produto “beleza\_saude” é a que vendeu mais em valor enquanto que a categoria “cama\_mesa\_banho” vendeu mais em quantidade.

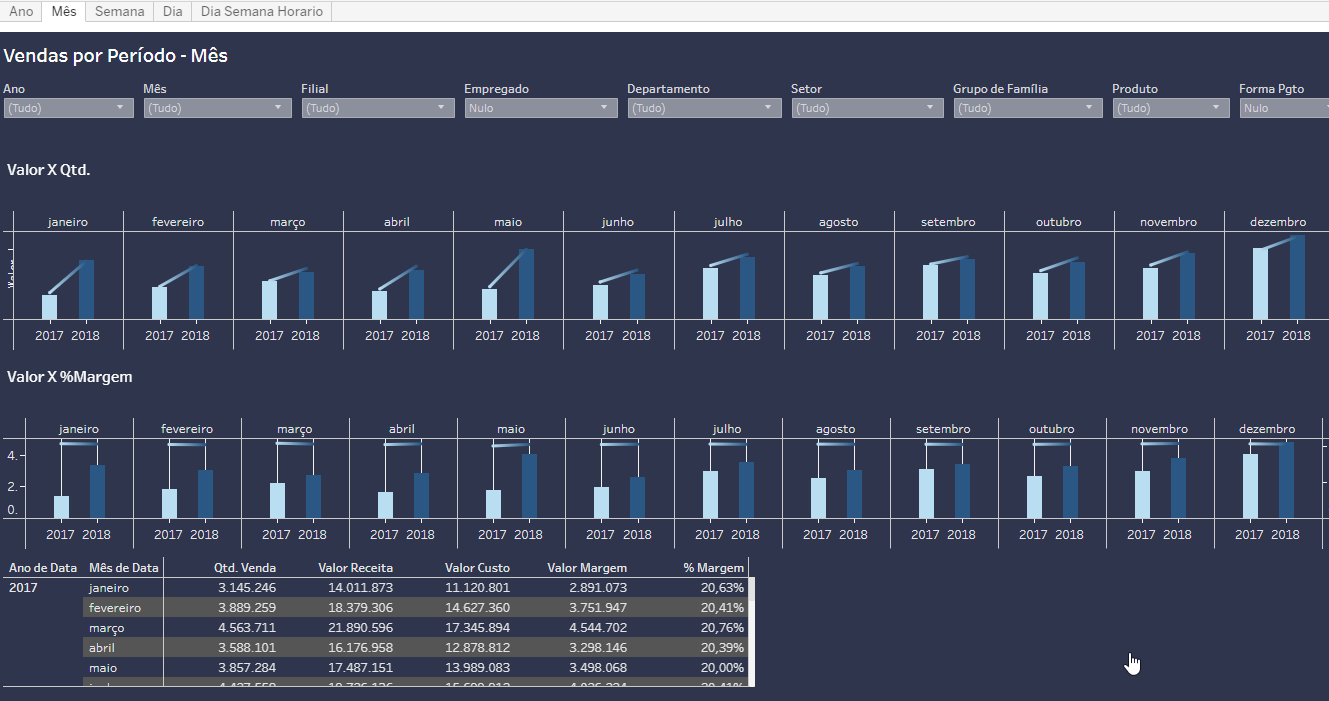


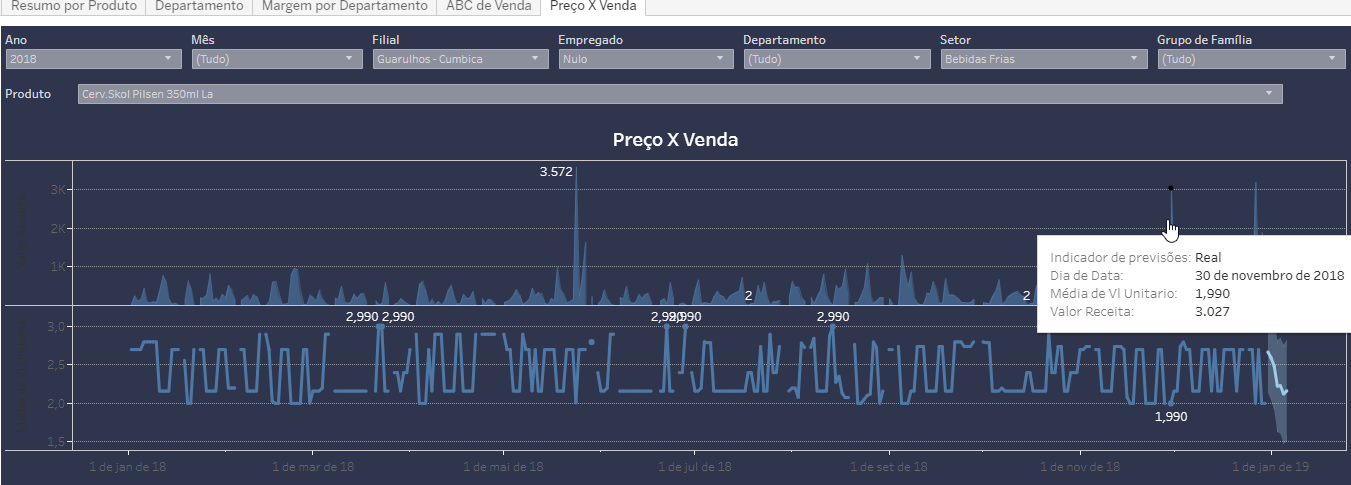
Grande pico de Vendas no dia 24/11/2017, dia do BlackFriday.

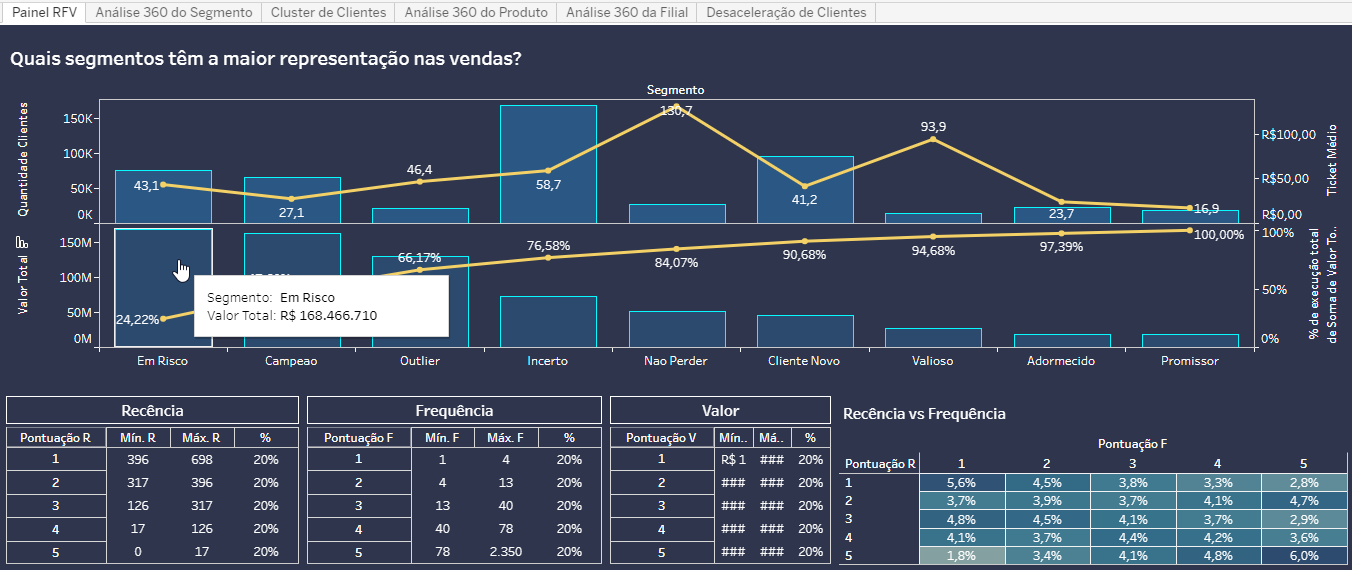
* Você poderia apresentar esses dados em um dashboard. Isso daria agilidade na tomada de decisão

Monteium projeto (Dashboards) no Tableau (vou tentar enviar o arquivo twbx).

Essa área (BI/Analytics) é a que tenho mais experiência e poderia desenvolver mais análises/dashboards. Se interessar tenho outro trabalho no tableau que ajudei a desenvolvercom outra fonte de dados e voltado para o varejo. Está bem mais bem elaboradoe levou meses de desenvolvimento.



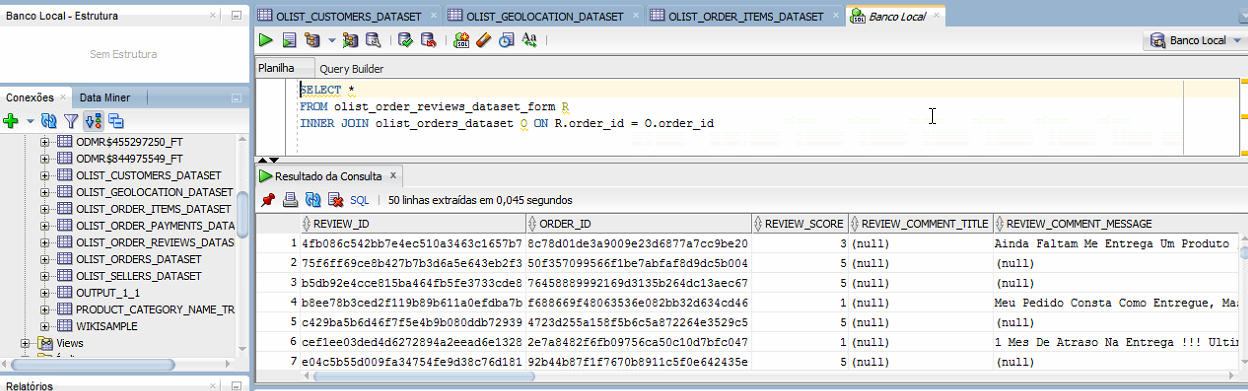




**Engenheiro(a) de dados**

* Gostaríamos de analisar suas habilidades com SQL, modelagem dimensional e integração de dados. Mostre seus conhecimento em processos de ETL e conceitos de Data Warehouse. Que tal replicar nossos datasets, remodelar em um banco de dados e apresentar as melhorias realizadas em sua criação.

Realizei Instalação Default do Banco de Dados Oracle (para utilizar o Oracle Data Miner) em uma máquina local Windows e fiz a importação das tabelas. Não sou DBA mas por trabalhar em vários projetos, sei que para BI é interessante aumentar o tamanho do bloco, criar índices para as colunas mais utilizadas e colher estatísticas.



* É possível utilizar o modelo proposto em um ambiente cloud? Quais plataformas ou serviços você utilizaria? Quais as vantagens do modelo escolhido em questões de performance?

Só tive contato com o ambiente Cloud da Oracle.

* Alguns membros do time dizem que a atual modelagem do banco de dados é adequada para o uso dos cientistas de dados e analistas de BI, porém, outros dizem que existem formas de modelar bancos de dados que trarão mais eficiência. Qual é a sua opinião sobre isso?

Acredito ser adequada. Apesar de não conhecer com detalhes, não senti dificuldade em trabalhar com as tabelas. Senti “falta” mas acredito que vocês tem informações mais ricas que não puderam compartilhar (por exemplo cadastro detalhado do cliente, renda, hierarquia de classificação de produtos).

* Estamos preocupados com o vertiginoso aumento do volume em nosso banco de dados atual? Você consideraria uma opção mais escalável ou devemos manter a estrutura existente?

Não tenho conhecimento suficiente para sugerir outra opção.

* Nossa ferramenta de visualização de dashboards está lenta e o nosso time detectou que o problema está na infraestrutura de dados. Como você abordaria esta situação do ponto vista de arquitetura de dados?
* Nosso banco de dados está hospedado na nuvem e nossas ferramentas de análise de dados são "on premisses". Você manteria este arranjo ou faria mudanças visando mais performance?

Na nossa empresa trabalhamos com o tableau server (em um ambiente próprio dentro do data center da Copel). Para deixar mais rápido, tiramos a conexão em tempo real e estamos trabalhando com o “extract” (fazendo atualizações incrementais toda manhã). Desenvolvemos no tableau desktop e publicamos para o server. A visualização no tableau server com “extract” ficou com desempenho muito bom.

* Nossa área operacional necessita de informações em tempo real, porém os diretores da empresa, que acompanham somente informações de KPIs mensais, alegam que isso é desnecessário e acarretaria custos. Qual é o seu posicionamento sobre isso?

Concordo que informação em tempo real acarretaria custos pois não podemos deixar lento o sistema de produção e soluções para isso são caras.Participei de um projeto na Rumo em que precisavam de informação em tempo real e foi adquirido pela empresa o Oracle Golden Gate para “Duplicar” as bases com integridade e velocidade.Não sei qual a informação que a área operacional precisa, mas o necessário o sistema ERP teria que fornecer.

Não sou Engenheiro de Dados, tenho alguma noção dos projetos no qual eu trabalhei.