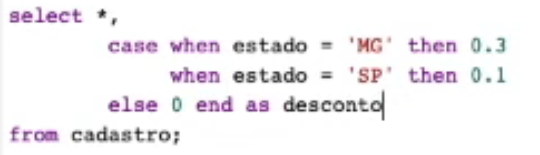
## **Semana 5: SQL, Polars e Discussões de Fóruns**

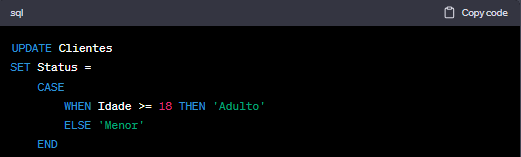
**SQL: Introdução, dicas e boas práticas**

1. Case When, o IF do SQL

É basicamente um IF para usar das mais variadas formas, em consultas, por exemplo, Queremos criar uma coluna na hora de fazer uma consulta a uma tabela em sql, essa coluna será o desconto aplicado a uma base de dados, sendo 30% a quem for de são paulo, 10% pra minas e 0 pro resto:

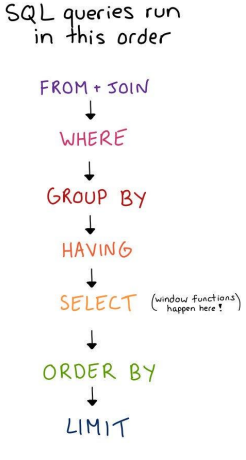


Essa coluna não ficará salva, apenas será usada na consulta, para salvar na tabela temos que fazer um update table:

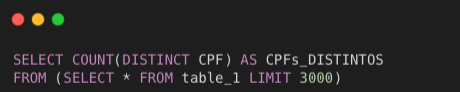


1. Ordem de execução em SQL

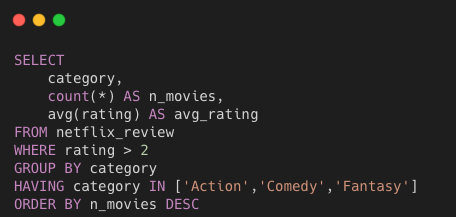
É importante saber a ordem em que os comando são executados em SQL, por exemplo, em um dataset de milhões de linhas, no qual você só quer alterar mil linhas, é mais eficiente fazer as alterações nas linhas já filtradas, para isso, precisamos saber a ordem do que acontece por trás do compilador.



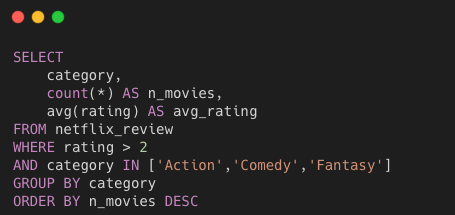
Por exemplo, se queremos colocar um limite de 3000 linhas para uma consulta rodar mais rápido, não podemos simplesmente chamar o LIMIT, uma vez que ele é executado após tudo ter acontecido, então não vai limitar a quantidade de linhas que foi tratada. Para isso, poderíamos simplesmente filtrar o FROM para puxar apenas 3000 linhas, fazendo com que a consulta se limite a essas linhas.



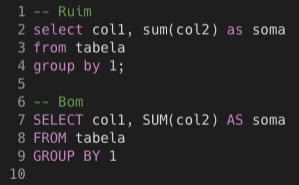
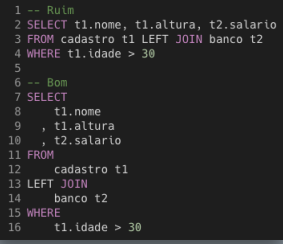
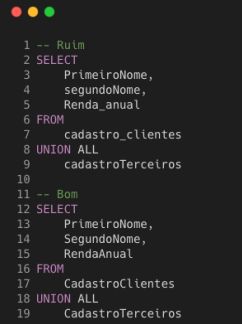
Outro exemplo em que a ordem de execução acarretaria em uma melhora na eficiência do código, é neste código abaixo:



Ele primeiro faz um GROUP BY e após ele, faz um filtro que poderia ser feito antes de GROUP BY, o que é muito ineficiente, pois poderíamos ter feito no WHERE, que é executado antes do GROUP BY, deixando o tempo de execução do GP mais baixo. Geralmente o HAVING é usado quando precisamos filtrar por algum valor dependente do GROUP BY, como soma, média, máximo, etc. A maneira certa de fazer esse filtro seria:



1. Boas Práticas em SQL

* Palavras Chaves e funções em letra maiúscula
* Código identado e alinhado
* Uso do mesmo padrão para o nome de variáveis, tabelas, schemas
* Evitar SELECT \*, é preferível especificar as colunas
* Evitar comentários desnecessários
* 
* 
* 

**Discussões de fóruns**

1. **Quando fazer o pré processamento e quando fazer o split?**

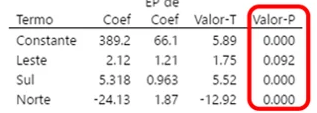
Alguns tipos de pré-processamentos, como tratamento de missing, scaler e encoder devem ser feitos após o split, uma vez que caso seja feito antes, estariam usando valores do teste para a construção do modelo. Por exemplo, quando vamos substituir os missing pela média antes de fazer o split, estaremos usando os valores dos dados de teste para fazer essa substituição, o que pode fazer com que o modelo tenha uma espécie de acesso aos dados de teste por conta disso. Porém, existem casos que o split não vai mudar o pré-processamento, por exemplo “eu preciso criar uma variável que utilize o número de dias entre a coluna 1 e a coluna 2”, independente de ser feito antes ou depois do split, o resultado será o mesmo, diferente de uma média, mediana, moda, por exemplo, que usaria os valores tanto de teste quanto treino, nos casos onde isso não acontece, podemos fazer o pré-processamento tanto antes quanto depois do split.

OBS: Mesmo depois de dividir em treino e teste, devemos substituir os missings do TESTE pela média calculada no TREINO

1. **P-valor em modelos de regressão linear múltipla**

Antes de compreender o que é o p-valor, precisamos entender o que é a hipótese nula(H0), sendo ela a hipótese inicial que temos em relação aos dados, geralmente tentamos desmentir ela, logo, a hipótese nula retrata que NÃO existe uma relação no fenômeno estudado, uma vez que queremos desmentir ela. De forma contrária, a hipótese alternativa é aquela que corrobora com o que estamos estudando, é ela que queremos legitimar ou apoiar. Enquanto geralmente a hipótese nula afirma não haver efeito, a alternativa diz existir um efeito. Enquanto isso, o p-valor é a probabilidade de se obter os resultados observados(são os dados coletados durante a pesquisa) considerando a H0 como sendo verdadeira, quanto menor o p-valor, mais evidências existem para descartar a H0.(geralmente esse limiar é 0.05).Basicamente é ele que vai dar a resposta para um teste de hipótese. Em regressão linear, os p-valor dos coeficientes de regressão linear verificam se aquele coeficiente é diferente de 0, uma vez que aqueles coeficientes encontrados podem não ser significantes, que tenham sido considerados melhores apenas por obra do acaso e que seus resultados não sejam diferentes de zero(sem efeito), é isso que o p-valor vai provar ou refutar. No caso da regressão linear, a H0 seria referente a não ter efeito, ou seja, não há diferença em ter aquele coeficiente ou não, como queremos refutar isso, quanto menor o p-valor, melhor, para esse caso

.Por outro lado, um valor-p maior (insignificante) sugere que as mudanças na preditora não estão associadas a mudanças na resposta, confirmando a H0.



Neste resultado, podemos ver que o termo leste está com o p-valor maior que o limiar 0.05, o que indica que ele não é estatisticamente significante para o modelo.