Preditiva.ai

Introdução a Extração e Visualização de Dados

Parte I

JOAO ESTEVAN LEONCIO DA SILVA BARBOSA - jestevan12@gmail.com - CPF: 134.982.877-70

Extração de Dados

O que é SQL?

Preditiva.ai

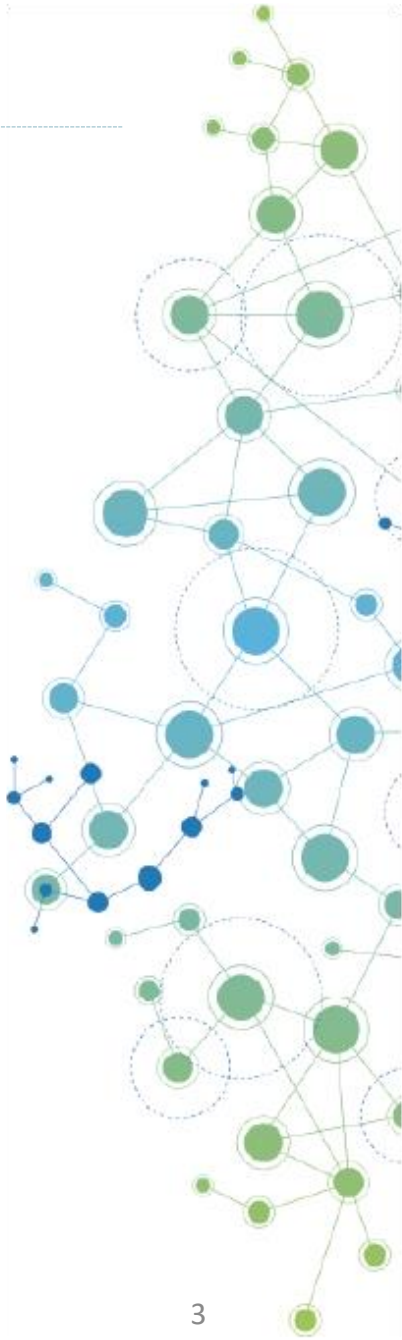
Agora que entendemos como os dados são organizados, é importante aprendermos como acessá-los, transformá-los e transportá-los para a nossa ferramenta de trabalho de forma eficiente.

Para atingirmos esse objetivo, utilizaremos a linguagem **SQL - Structured Query Language**, ou Linguagem de Consulta Estruturada.

Inicialmente desenvolvida pela IBM no início dos anos 1970, ela ainda é muito utilizada atualmente devido a sua versatilidade, facilidade de uso e compatibilidade com diversos tipos de bancos de dados.

Entre as principais características do SQL podemos citar

- Fácil compreensão
- Acesso direto aos dados
- Fácil de replicar e auditar
- Capacidade de acessar diversas tabelas simultaneamente
- Capacidade de trabalhar com bilhões de registros



Introdução ao SQL:

Seleção e Filtro de Variáveis

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

Nas nossos experimentos vamos utilizar o famoso conjunto de dados Titanic.

Fonte: https://aventurasnahistoria.uol.com.br/noticias/reportagem/neste-dia-em-1912-o-rms-titanic-colidia-com-um-iceberg-no-mais-impactante-desastre-do-seculo-20.phtml

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

Neste conjunto de dados temos diversas informações sobre os passageiros, conforme o metadados apresentado abaixo:

Campo	Descrição	Tipo Campo	Tipo Variável
PassengerId	Código único de identificação do passageiro	Numérico	Qualitativa Nominal
Survived	Indicador de sobrevivência: 1 - sobreviveu, 0 - não sobreviveu	Numérico	Qualitativa Nominal
Pclass	Classe das acomodações no navio: 1, 2 ou 3	Numérico	Qualitativa Ordinal
Name	Nome do passageiro	Texto	Qualitativa Nominal
Sex	Gênero do passageiro	Texto	Qualitativa Nominal
Age	Idade do passageiro	Numérico	Quantitativa Discreta
SibSp	Número de irmãos ou esposa a bordo	Numérico	Quantitativa Discreta
Parch	Número de pais ou filhos a bordo	Numérico	Quantitativa Discreta
Ticket	Número do ticket	Texto	Qualitativa Nominal
Fare	Valor da tarifa	Numérico	Quantitativa Contínua
Cabin	Código da cabine	Texto	Qualitativa Nominal
Embarked	Porto de embarque: **C** - Cherbourg, **Q** - Queenstown, **S** - Southampton	Texto	Qualitativa Nominal

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

Para iniciar a exploração de uma tabela, podemos selecionar uma amostra das primeiras observações com todas as variáveis.

SELECT: indica quais variáveis serão selecionadas. Para selecionar todas você pode utilizar o '*'.

```
SELECT TOP 10
    *
FROM
    titanic
```

TOP **n**: seleciona apenas as primeiras **n** observações. Útil nas explorações iniciais.

FROM: indica a origem dos dados.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
	SMALLINT	BIT	TINYINT	NVARCHAR (100)	NVARCHAR (50)	FLOAT	TINYINT	TINYINT	NVARCHAR (50)	FLOAT	NVARCHAR (50)	NVARCHAR (50)
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	725	NULL	S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	712.833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	NULL	S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	531	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	805	NULL	S
6	6	0	3	Moran, Mr. James	male	NULL	0	0	330877	84.583	NULL	Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	518.625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	NULL	S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	111.333	NULL	S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	300.708	NULL	C

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

Podemos perceber que a variável **Cabin** apresenta muitas observações com **NULL**. Essa identificação representa, no linguajar de análise de dados, um ***missing value***, ou **informação faltante**.
Vamos refazer a query **filtrando** apenas os registros que não possuam valores faltantes na variável **Cabin**.

```
SELECT TOP 10
   *
FROM
   titanic
WHERE
   Cabin is NOT NULL
```

WHERE: filtra as observações baseado em condições

	PassengerId	Survived	Pc		Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
	SMALLINT	BIT	TINYINT	NVARCHAR (100)	NVARCHAR (50)	FLOAT	TINYINT	TINYINT	NVARCHAR (50)	FLOAT	NVARCHAR (50)	NVARCHAR (50)
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	712.833	C85	C
2	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	531	C123	S
3	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	518.625	E46	S
4	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	167	G6	S
5	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	2.655	C103	S
6	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
7	24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	355	A6	S
8	28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C27	S
9	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NULL	1	0	PC 17569	1.465.208	B78	C
10	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49	1	0	PC 17572	767.292	D33	C

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

Ainda estamos apenas no início, mas para evoluirmos com organização é importante seguir alguns padrões de formatação das consultas.

Esses padrões **facilitam** a **utilização** e a **manutenção** das consultas.

```
SELECT TOP 10
   *
FROM
   titanic
WHERE
   Cabin is NOT NULL
```

Formatação

1. **Palavras-chave**: Sempre em letras **maiúsculas**.
2. **Nome de campos e tabelas**: Sempre em letras **minúsculas**.
3. **Espaços em branco**: Evitar usar em nomes de campos e nomes de tabelas.
4. **Endentação**: Utilize para facilitar a leitura e entendimento, e a manutenção.

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

A seguir vamos nos aprofundar em algumas outras funcionalidades que podem ser muito úteis no nosso dia a dia de análise de dados.

- **WHERE**: Filtra os registros de acordo com as condições especificadas
- **ORDER BY**: Ordena o resultado

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

WHERE: FILTRAR as observações

A palavra-chave **WHERE** é muito importante por permitir que eliminemos do resultado da consulta as observações não desejadas. Com isso, já realizamos uma primeira etapa de preparação dos dados para a análise.

Operadores lógicos: Podem ser utilizados em campos numéricos ou texto. Nos campos texto é obedecida a ordem alfabética.

Operador	Exemplo Numérico	Exemplo Texto	Descrição
=	WHERE Age = 18	WHERE Embaked = 'C'	igual a
!=	WHERE Age != 18	WHERE Embaked != 'C'	diferente de
>	WHERE Age > 18	WHERE Embaked > 'C'	maior do que
>=	WHERE Age >= 18	WHERE Embaked >= 'C'	maior ou igual a
<	WHERE Age < 18	WHERE Embaked < 'C'	menor do que
<=	WHERE Age <= 18	WHERE Embaked <= 'C'	menor ou igual a

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

WHERE: FILTRAR as observações

Em campos texto o **LIKE** é o operador que nos permite criar condições mais sofisticadas. Ou seja, podemos filtrar por parte do texto, seja ele o começo, meio ou fim.

Exemplo	Descrição
WHERE Cabin LIKE 'C%'	Código da cabine começa com 'C'
WHERE Ticket LIKE '%7'	Código do ticket termina com '7'
WHERE Name LIKE '%, Dr. %'	Nome do passageiro contém ', Dr. '

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

WHERE: FILTRAR as observações

Quando temos mais de um valor, numérico ou texto, que queremos utilizar como filtro, podemos utilizar o operador **IN**.

Exemplo	Descrição
WHERE Embarked IN ('C','Q')	Embarque realizado nos portos Cherbourg ou Queenstown
WHERE Pclass IN (2,3)	Cabines de 2ª ou 3ª classe

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

WHERE: FILTRAR as observações

Para selecionar um intervalo numérico, um outro operador importante é o **BETWEEN**. Ao invés de utilizar:

```
WHERE Age >= 18 AND Age <= 25
```

Podemos utilizar:

```
WHERE Age BETWEEN 18 AND 25
```

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

WHERE: FILTRAR as observações

Podemos utilizar os operadores **AND** e **OR** para combinar diferentes condições e construir um critério de seleção mais complexo:

Idade maior ou igual a 18 do sexo feminino:

```
WHERE Age >= 18 AND sex = 'female'
```

Idade maior ou igual a 18 OU sexo feminino:

```
WHERE Age >= 18 OR sex = 'female'
```

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

WHERE: FILTRAR as observações

Nas situações em que o critério de exclusão é mais simples do que o de inclusão no resultado, podemos utilizar o operador **NOT**. Ele inverte o efeito lógico dos operadores utilizados.

Exemplo	Descrição
WHERE nome NOT LIKE '%, Dr. %'	Nome NÃO contém ', Dr. '
WHERE Embarked NOT IN ('C','Q')	Embarque NÃO realizado nos portos Cherbourg ou Queenstown
WHERE Age NOT BETWEEN 18 and 25	Idade não está entre 18 e 25

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

ORDER BY: Ordena o resultado da consulta

Caso desejemos ordenar o resultado da consulta, utilizamos o comando **ORDER BY**.
Também é possível ordenar de forma inversa ou decrescente. Nesse caso, utilizamos o ORDER BY seguido do nome do campo e a opção **DESC**.

```
SELECT TOP 10
   *
FROM
   titanic
ORDER BY
   PassengerId DESC
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ti
	SMALLINT	BIT	TINYINT	NVARCHAR (100)	NVARCHAR (50)	FLOAT	TINYINT	TINYINT	NVARCH
1	891	0	3	Dooley, Mr. Patrick	male	32	0	0	370376
2	890	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369
3	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NULL	1	2	W./C. 66
4	888	1	1	Graham, Miss. Margaret Edith	female	19	0	0	112053
5	887	0	2	Montvila, Rev. Juozas	male	27	0	0	211536
6	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39	0	5	382652
7	885	0	3	Sutehall, Mr. Henry Jr	male	25	0	0	SOTON/C
8	884	0	2	Banfield, Mr. Frederick James	male	28	0	0	C.A./SOT
9	883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22	0	0	7552
10	882	0	3	Markun, Mr. Johann	male	33	0	0	349257

Extração de Dados

Introdução ao SQL: Seleção e Filtro de Variáveis

Preditiva.ai

Foi pedido um dashboard com a lista e perfil dos tripulantes da primeira classe, considerando as variáveis: Sexo, Idade e Classe.

Para selecionar apenas as variáveis de nosso interesse, listamos elas no **SELECT** separado por **vírgulas**. Essas variáveis serão **as visualizadas** na nossa query.

```
SELECT
    PassengerId,
    Name,
    Sex,
    Age
FROM
    titanic
WHERE
    Pclass = 1
```

Separe com vírgulas cada variável

A última variável listada não tem vírgula

	PassengerId	Name	Sex	Age
1	2	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
2	4	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
3	7	McCarthy, Mr. Timothy J	male	54
4	12	Bonnell, Miss. Elizabeth	female	58
5	24	Sloper, Mr. William Thompson	male	28
6	28	Fortune, Mr. Charles Alexander	male	19
7	31	Uruchurtu, Don. Manuel E	male	40
8	32	Spencer, Mrs. William Augustus (Marie Eugenie)	female	[NULL]
9	35	Meyer, Mr. Edgar Joseph	male	28
10	36	Holverson, Mr. Alexander Oskar	male	42
11	53	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49
12	55	Ostby, Mr. Engelhart Cornelius	male	65
13	56	Woolner, Mr. Hugh	male	[NULL]
14	62	Icard, Miss. Amelie	female	38
15	63	Harris, Mr. Henry Birkhardt	male	45

Preditiva.ai