 **📚 Lesson 03**  | Filtering with WHERE

[WHERE](#cr72lch9nj73)

[BETWEEN & IN](#80wfxzv55bl1)

[LIKE](#p7dov048v1dn)

[IS NULL & IS NOT NULL](#b1o8lvurm9z0)

[AND, OR, & NOT](#1eykl1wjn5x4)

[COMBINING AND, OR & NOT](#ym2y8pmd42r2)

 **WHERE Practice**

— Filtering Data

PROMPT: Let’s practice using the WHERE keyword by performing some basic filtering on the **sales\_data\_cars table**, which comprises a sample of 2500 car sales in two regions of California.

1. Write a query that returns all cars with a highway fuel mileage of 25 mpg or less. How many cars have this property?

|  |
| --- |
| select count(\*)  from sales\_data\_cars  where fuel\_economy\_highway <= 25 |

|  |
| --- |
| 357 |

1. Modify the query to answer this question: What is the oldest car with a highway fuel mileage of 25 or less?

|  |
| --- |
| select min(year)  from sales\_data\_cars  where fuel\_economy\_highway <= 25 |

|  |
| --- |
| 2005 |

1. Write a query that returns the make, model, year, and both city and highway fuel economy for cars manufactured by Ford. How many vehicles in the dataset were manufactured by Ford?

|  |
| --- |
| select  make,  model,  year,  fuel\_economy\_city,  fuel\_economy\_highway  from sales\_data\_cars  where make = 'Ford' |

|  |
| --- |
| 118 |

1. Modify the query to answer this question: What is the model of Ford car that has the highest city fuel mileage?

|  |
| --- |
| select  make,  model,  year,  fuel\_economy\_city,  fuel\_economy\_highway  from sales\_data\_cars  where make = 'Ford'  order by 4 DESC |

|  |
| --- |
| Fusion |

 **BETWEEN and IN Practice**

— Filtering Data

PROMPT: You’ve just learned some additional keywords to help you perform filtering, BETWEEN and IN. Let’s practice using them by looking at data from house sales in Ames, Iowa (**iowa\_ames\_housing**).

There are a lot of columns in the data table, so make sure you select a relevant subset of columns in your SELECT statements so that the output is easier to read. Write a query that returns columns: saleprice, lot\_area, neighborhood, bldg\_type, and house\_style. You might want to save this base query, since we’ll be coming back to this dataset later in the module.

1. Warm up: What is the highest price that a house was sold for in the data, and in which neighborhood is it located?

|  |
| --- |
| select  saleprice,  lot\_area,  neighborhood,  bldg\_type,  house\_style  from iowa\_ames\_housing  order by 1 desc |

|  |
| --- |
| 625000 in NoRidge |

1. Modify your query and use the app interface to answer this question: How many houses in the dataset were sold for between $200 000 and $400 000?

|  |
| --- |
| select  saleprice,  lot\_area,  neighborhood,  bldg\_type,  house\_style  from iowa\_ames\_housing  Where saleprice between 200000 and 400000 |

|  |
| --- |
| 812 |

1. Modify your query one more time to answer this question: What is the largest lot size for a townhouse (Twnhs or TwnhsE) sold in the dataset?

|  |
| --- |
| select  saleprice,  lot\_area,  neighborhood,  bldg\_type,  house\_style  from iowa\_ames\_housing  Where bldg\_type in ('Twnhs', 'TwnhsE')  order by 2 desc |

|  |
| --- |
| 14963 |

 **LIKE Practice**

— Filtering Data

PROMPT: Text data can often be tricky to work with, but the LIKE and ILIKE keywords can be very effective towards getting information out of text fields. Let’s see these keywords in action by looking at data from the information about selected businesses in the Yelp dataset (**yelp\_business**).

1. Warm up: How many restaurants in our dataset are in a city named “Saint Joseph”?

|  |
| --- |
| select count(\*)  from yelp\_business  where city in ('Saint Joseph') |

|  |
| --- |
| 8 |

1. There’s a lot of cities in the dataset that start with “Saint”. How many restaurants are located in a city that starts with “Saint”?

|  |
| --- |
| select count(\*)  from yelp\_business  where city like 'Saint%' |

|  |
| --- |
| 427 |

1. How many restaurants are listed as being in “Gilbert”? It’s worth noting that the city name has been entered into the database as “GILBERT” as well as “Gilbert”.

|  |
| --- |
| select count(\*)  from yelp\_business  where city in ('Gilbert', 'GILBERT') |

|  |
| --- |
| 618 |

1. How many restaurants list “African” food among their categories? Note that this tag can appear anywhere in the categories field.

|  |
| --- |
| select count(\*)  from yelp\_business  WHERE categories like '%African%' |

|  |
| --- |
| 165 |

 **IS NULL Practice**

— Filtering Data

PROMPT: While a null value represents missing data, being able to detect nulls can lead to other insights into the data. Let’s see this in action by practicing the IS NULL keyword with the IMDB dataset (**imdb\_data**), which has basic information on about 5 000 movies and shows.

1. How many movies in the dataset have missing values for their durations?

|  |
| --- |
| select \*  from imdb\_data  where duration is null |

|  |
| --- |
| 14 |

1. Of those movies, which one has the highest IMDB rating?

|  |
| --- |
| select \*  from imdb\_data  where duration is null  order by imdb\_score desc  limit 1 |

|  |
| --- |
| War and Peace |

1. Run two queries to answer this question: Is it more common in our dataset for a movie to be missing its budget, or for it to be missing its gross?

|  |
| --- |
| select count(\*)  from imdb\_data  where budget is null  select count(\*)  from imdb\_data  where gross is null |

|  |
| --- |
| 491 vs 883 null gross more common |

 **AND, OR & NOT Practice**

— Filtering Data

PROMPT: Logical operators allow us to apply multiple filter conditions to our data. Practice using them by going back to the Ames, Iowa house sales dataset (**iowa\_ames\_housing**).

As a reminder, to reduce the amount of data output, make sure your queries only SELECT for the columns saleprice, lot\_area, neighborhood, bldg\_type, and house\_style.

1. How many houses are in the dataset with a sale price over $300 000 or a lot size greater than 15 000 square feet?

|  |
| --- |
| select  saleprice,  lot\_area,  neighborhood,  bldg\_type,  house\_style  from iowa\_ames\_housing  where saleprice > 300000 or lot\_area > 15000 |

|  |
| --- |
| 422 |

1. How many sales of single-family homes (1Fam) with a lot size of less than 10 000 square feet are there in the dataset?

|  |
| --- |
| select  saleprice,  lot\_area,  neighborhood,  bldg\_type,  house\_style  from iowa\_ames\_housing  where bldg\_type = '1Fam' and lot\_area < 10000 |

|  |
| --- |
| 1270 |

1. Of those houses, what is the lowest sale price?

|  |
| --- |
| select  saleprice,  lot\_area,  neighborhood,  bldg\_type,  house\_style  from iowa\_ames\_housing  where bldg\_type = '1Fam' and lot\_area < 10000  order by saleprice |

|  |
| --- |
| 12789 |

1. What is the smallest lot size for a building that IS NOT a townhouse (Twnhs or TwnhsE)?

|  |
| --- |
| select  saleprice,  lot\_area,  neighborhood,  bldg\_type,  house\_style  from iowa\_ames\_housing  where bldg\_type not in ('TwnhsE','Twnhs')  order by lot\_area |

|  |
| --- |
| 2500 |

 **Combining AND, OR, & NOT Practice**

— Filtering Data

PROMPT: Let’s write some example queries to show the importance of writing clearly-defined code. In this practice, we’ll return to the Yelp businesses dataset (**yelp\_business**).

1. Warm up: How many restaurants in our dataset end with the possessive ‘s (e.g. Wendy’s, McDonald’s)? You can match on an apostrophe in a string by typing it two times in a row.

|  |
| --- |
| select count(\*)  from yelp\_business  where name like '%''s' |

|  |
| --- |
| 3255 |

1. How many of those businesses DO NOT include Fast Food among their categories tags?

|  |
| --- |
| select count(\*)  from yelp\_business  where name like '%''s' and categories not like '%Fast Food%' |

|  |
| --- |
| 1432 |

1. How many businesses in Nevada (NV) have a Burgers or Pizza category tag? HINT: Make sure you look at your output to make sure you have the filtering done correctly!

|  |
| --- |
| select count(\*)  from yelp\_business  where state='NV' and (categories like '%Pizza%' or categories like '%Burgers%') |

|  |
| --- |
| 1715 |

1. Challenge: How many of those businesses are NOT in Henderson, Las Vegas, or North Las Vegas? You might need to iterate on your query multiple times to clean everything up!

If you just added city NOT IN ('Henderson', 'Las Vegas', 'North Las Vegas') to your query, you would have gotten 30 restaurants back, but should have seen some leftover rows with 'Vegas' in their city names. In order to automatically filter them out, we need to create a more elaborate condition.

|  |
| --- |
| SELECT  name,  city,  state,  categories  FROM yelp\_business  WHERE  state = 'NV'  and (  categories like '%Burgers%'  or categories like '%Pizza%'  )  and (  city != 'Henderson'  AND city NOT LIKE '%Vegas%'  ) |

|  |
| --- |
| 24 |