Airfare SQL Challenge

Link to challenge on Codecademy
airefare_data.csv
Platform - To do this project, I am using DB Browser for SQLite.

Column Descriptions

Column Name	Description	Type
Year	Year	Number
quarter	Quarter	Number
citymarketid_1	City market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market	Number
citymarketid_2	City market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market	Number
city1	City1 is used to consolidate airports serving the same city market	Plain Text
city2	City2 is used to consolidate airports serving the same city market	Plain Text
nsmiles	Non-Stop market miles (using radian measure)	Number
passengers	Passenger per day	Number
fare	Overall average fare	Number
carrier_lg	Carrier with the largest market share	Plain Text
large_ms	Market share for the carrier with the largest market share	Number
fare_lg	Average fare for the carrier with the largest market share	Number

carrier_low	Carrier with the lowest fare	Plain Text
lf_ms	Market share for the carrier with the lowest fare	Number
fare_low	Average fare for the carrier with the lowest fare	Number
table_1_flag	Flag for Table 1 subset. Top 1,000 Contiguous State City Pair Markets.	Number
Geocoded_City1	Geocoded - City1 is used to consolidate airports serving the same city market with Latitude and Longitude	~~Point~~ Text
Geocoded_City2	Geocoded - City2 is used to consolidate airports serving the same city market with Latitude and Longitude	~~Point~~ Text

Data Exploration - Familiarize yourself with the dataset.

What range of years are represented in the data?

```
SELECT MIN(year), MAX(year)
FROM airfare_data;
```

output:

	MIN(year)	MAX(year)
1	1996	2018

What are the shortest and longest-distanced flights, and between which 2 cities are they?

```
SELECT city1, city2, nsmiles
FROM airfare_data
where nsmiles = (select max(nsmiles) from airfare_data) or nsmiles = (select
min(nsmiles) from airfare_data)
```

Hmm, this gives me the correct answers (flights between Miami and Seattle as the max, and flights between LA and San Diego as the min), but gives me unnecessary duplicates. I'll need to select distinct.

```
SELECT distinct city1, city2, nsmiles
FROM airfare_data
where nsmiles = (select max(nsmiles) from airfare_data) or nsmiles = (select
min(nsmiles) from airfare_data)
```

output:

	city1	city2	nsmiles
1	Miami, FL (Metropolitan Area)	Seattle, WA	2724
2	Los Angeles, CA (Metropolitan Area)	San Diego, CA	109

How many distinct cities are represented in the data (regardless of whether it is the source or destination)?

Hint: We can use UNION to help fetch data from both the city1 and city2 columns. Note the distinction between UNION and UNION ALL.

Considering I don't want duplicate values, I'll use UNION.

```
SELECT DISTINCT city1 FROM airfare_data
UNION
SELECT DISTINCT city2 FROM airfare_data
```

Output preview (163 rows):

	city1
1	Albany, NY
2	Albuquerque, NM
3	Allentown/Bethlehem/Easton, PA
4	Amarillo, TX
5	Appleton, WI
6	Asheville, NC

Analysis - Further explore and analyze the data

Which airline appear most frequently as the carrier with the lowest fare (ie. carrier_low)? How about the airline with the largest market share (ie. carrier_lg)?

Starting with the carrier with the lowest fare

```
SELECT carrier_low, count(*)
FROM airfare_data
GROUP by carrier_low
ORDER by 2 desc
```

output preview (54 rows):

	carrier_low	count(*)
1	WN	29652
2	DL	8369
3	AA	7313
4	US	6527
5	FL	5997
6	F9	4382

It looks like airline WN (which is SouthWest Airlines according to a quick Google search) is most frequently (by a wide margin) the carrier with the lowest fare.
If I wanted only the winning row, I could include a LIMIT 1 clause.

Most frequently largest market share:

```
SELECT carrier_lg, count(*)
FROM airfare_data
GROUP by carrier_lg
ORDER by 2 desc
```

output preview (43 rows):

	carrier_lg	count(*)
1	WN	23659
2	DL	15789
3	AA	11375
4	US	8949
5	UA	8090
6	NW	5307

WN (SouthWest) is also the winner when it comes to largest market share frequency.

How many instances are there where the carrier with the largest market share is not the carrier with the lowest fare? What is the average difference in fare?

I'll start with instances where the carrier with the largest market share is not the carrier with the lowest fare.

```
SELECT * FROM airfare_data
where carrier_lg != carrier_low
```

gives me all of those instances. output preview (59851 rows):

	Year	quarter	citymarketid_1	citymarketid_2	city1	city2	nsmiles	pass
1	2000	4	30397	33198	Atlanta, GA (Metropolitan Area)	Kansas City, MO	692	
2	2007	4	32575	34614	Los Angeles, CA (Metropolitan Area)	Salt Lake City, UT	590	
3	2004	4	32337	31650	Indianapolis, IN	Minneapolis/St. Paul, MN	503	
4	2008	4	30194	30559	Dallas/Fort Worth, TX	Seattle, WA	1670	
5	2010	1	32575	33244	Los Angeles, CA (Metropolitan Area)	Memphis, TN	1619	
6	1996	3	31057	30198	Charlotte, NC	Pittsburgh, PA	366	

The number of rows tells us the number of instances this occurs. But if we want an actual output with that answer, I'll use a subquery with the previous query to get a count.

```
WITH instances as (SELECT * FROM airfare_data
where carrier_lg != carrier_low)
SELECT count(*) from instances
```

output:

	count(*)
1	59851

I'll use that same subquery to answer the second part of the prompt, finding the average differences in fares between the airline with the largest market share and that with the lowest fare.

```
WITH instances as (SELECT * FROM airfare_data
where carrier_lg != carrier_low)

SELECT avg(fare_lg - fare_low) from instances
```

output:

avg(fare_lg - fare_low)
49.4636478922666

I can also round that answer to be more reader friendly

```
WITH instances as (SELECT * FROM airfare_data
where carrier_lg != carrier_low)

SELECT round(avg(fare_lg - fare_low),2) as "Average difference in fares"
from instances
```

output:

	Average difference in fares
1	49.46

Additional Challenges

What is the percent change in average fare from 2007 to 2017 by flight? How about from 1997 to 2017?

Hint: We can use the WITH clause to create temporary tables containing the airfares, then join them together to compare the change over time.

First I'll look for the average fare in 1997, 2007 and 2017

```
SELECT year, avg(fare) as avg_fare
FROM airfare_data
GROUP by Year
HAVING year = 1997 or year = 2007 or year = 2017
```

output:

	Year	avg_fare
1	1997	176.737056471764
2	2007	183.121925
3	2017	218.3373575

To calculate Percent Change :

$$Percent\ Change = \frac{Change}{Original} \times 100$$

So to find the change, I'll need to subtract the two fares.

```
WITH fares as (SELECT year, avg(fare) as avg_fare
FROM airfare_data
GROUP by Year
HAVING year = 1997 or year = 2007 or year = 2017)

SELECT ((SELECT avg_fare FROM fares WHERE year = 2017) - (SELECT avg_fare FROM fares
WHERE year = 2007))
```

Output:

	es WHERE year = 2017) - (SELECT avg_fare FROM
1	35.2154325000006

Then divide by the 2007 fare, and multiply by 100 to convert the decimal into percent form.

```
WITH fares as (SELECT year, avg(fare) as avg_fare
FROM airfare_data
GROUP by Year
HAVING year = 1997 or year = 2007 or year = 2017)

SELECT ((SELECT avg_fare FROM fares WHERE year = 2017) - (SELECT avg_fare FROM fares
WHERE year = 2007))*100
/(SELECT avg_fare FROM fares WHERE year = 2007) as Percent_change
```

Output:

	Percent_change
1	19.2305932236135

Thus there was an approximately 19.23% increase in average fare from 2007 to 2017.

Now to do the same for the percent change between 1997 to 2017.

```
FROM airfare_data
GROUP by Year
HAVING year = 1997 or year = 2007 or year = 2017)
```

```
SELECT ((SELECT avg_fare FROM fares WHERE year = 2017) - (SELECT avg_fare FROM fares
WHERE year = 1997))*100
/(SELECT avg_fare FROM fares WHERE year = 1997) as Percent_change
```

Output:

	Percent_change
1	23.5379619072032

Thus there was an approximately 23.54% increase in average fare from 1997 to 2017.

How would you describe the overall trend in airfares from 1997 to 2017, as compared 2007 to 2017?

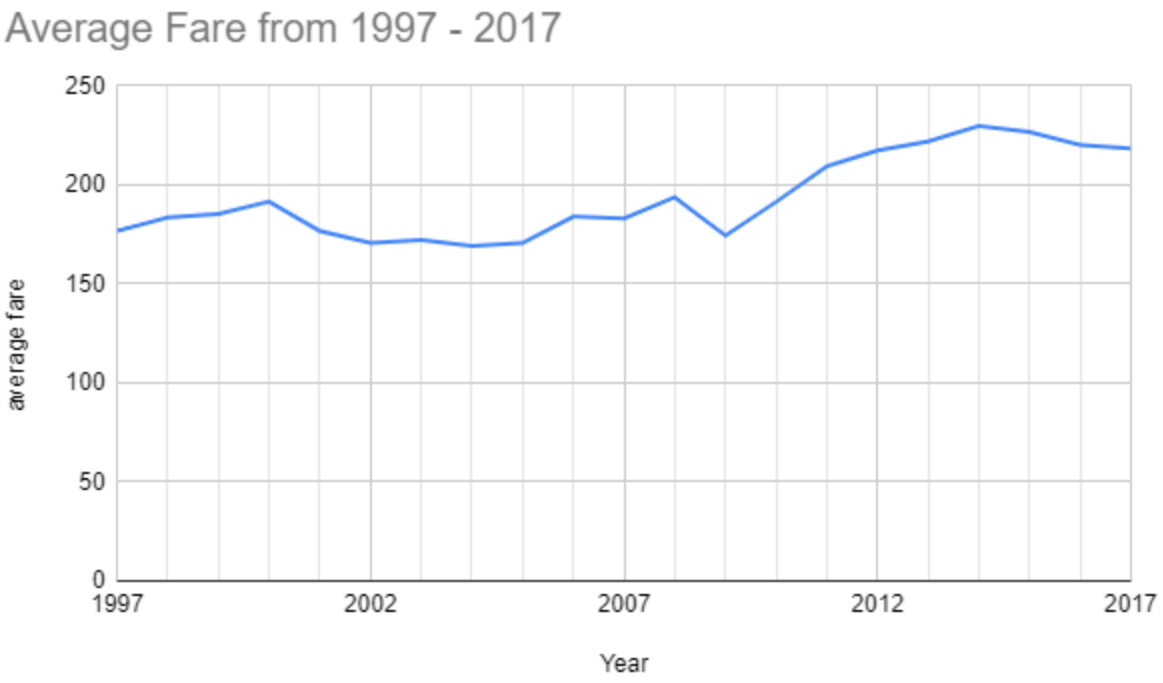
To look at the average fare for all years from 1997 - 2017

```
SELECT year, round(avg(fare),2) as avg_fare
FROM airfare_data
GROUP by Year
HAVING year BETWEEN 1997 and 2017
order by year
```

output:

Year	avg_fare
1997	176.74
1998	183.63
1999	185.31
2000	191.57
2001	176.66
2002	170.57
2003	172.04
2004	169.01
2005	170.58
2006	183.93
2007	183.12
2008	193.67
2009	174.43
2010	191.36
2011	209.36
2012	217.26
2013	221.95
2014	229.6
2015	226.6
2016	220.22
2017	218.34

Throwing this into google sheets for a super quick visualization:



Between 1997 and 2007, the average fare was overall pretty consistent. Prices fluctuated a bit, but leveled out in the long term. From 2007 to 2017 however, prices were overall trending upwards to steadily climb.

What is the average fare for each quarter? Which quarter of the year has the highest overall average fare? lowest?

```
SELECT round(avg(fare),2) as avg_fare, quarter
FROM airfare_data
GROUP by quarter
```

Output:

	avg_fare	quarter
1	195.79	1
2	195.03	2
3	191.31	3
4	190.44	4

Quarter 1 has the highest average fare, and quarter 4 has the lowest.
If I wanted to answer these more directly with specific queries, I could either do subqueries for the min/max avg_fare, or I could just add an ORDER BY avg_fare (ascending for the lowest/ descending for the highest) statement, and LIMIT 1.