Introduction to SQL
Understanding HAVING versus WHERE

The Basics

The key difference between HAVING and WHERE is the level at which filtering takes place.

- Filtering from WHERE takes place *before* aggregation occurs and is applied at the base record level.
- Filtering from HAVING takes place *after* aggregation and is applied to the aggregated data.

Sample Data

For this example we'll use a set of dummy data from a company specializing in winter hiking gear. Our table is called Raw_Invoices and includes the following row:

InvoiceDate	InvoiceNumber	Brand	ItemDescription	Qty	PaidAmt
10/1/2020	362559	COATS	Mountaineer HD	1	227.45
10/1/2020	302225	SOCKS	Thermal Toasty Toes	1	22.99
10/2/2020	362561	SCARVES	Cozy Cravat (electric)	1	110.00
10/3/2020	483277	PANTS	Deluxe Polar Pants	1	138.55
10/3/2020	362563	GLOVES	Condor Shur-grip	1	52.15
10/4/2020	483278	ACCESORY	Window decal	1	3.50
10/5/2020	362564	SCARVES	Cozy Cravat (electric)	1	110.00
10/6/2020	302226	PANTS	Deluxe Polar Pants	2	277.10
10/6/2020	362566	SCARVES	Cozy Cravat (electric)	1	110.00
10/7/2020	483279	SOCKS	Thermal Toasty Toes	2	45.98
10/9/2020	362568	SOCKS	Thermal Toasty Toes	1	22.99
10/9/2020	362569	PANTS	Deluxe Polar Pants	1	138.55
10/9/2020	483280	COATS	Mountaineer HD	1	227.45
10/10/2020	362571	GLOVES	Condor Shur-grip	1	52.15
10/10/2020	302227	ACCESORY	Window decal	1	3.50
10/11/2020	362572	SOCKS	Thermal Toasty Toes	1	22.99
10/12/2020	362573	COATS	Mountaineer HD	1	227.45
10/13/2020	362574	SCARVES	Cozy Cravat (electric)	1	110.00
10/14/2020	362575	GLOVES	Condor Shur-grip	2	104.30
10/14/2020	362576	COATS	Mountaineer HD	1	227.45
10/14/2020	483281	ACCESORY	Window decal	1	3.50
10/17/2020	362577	GLOVES	Condor Shur-grip	1	52.15
10/17/2020	362578	GLOVES	Condor Shur-grip	1	52.15
10/18/2020	362579	COATS	Mountaineer HD	1	227.45
10/22/2020	362580	SOCKS	Thermal Toasty Toes	2	45.98
10/23/2020	483282	ACCESORY	Window decal	1	3.50
10/24/2020	302229	GLOVES	Condor Shur-grip	2	104.30
10/24/2020	302230	ACCESORY	Window decal	1	3.50
10/27/2020	362582	COATS	Mountaineer HD	1	227.45
10/30/2020	362583	GLOVES	Condor Shur-grip	2	104.30

The Business Question

Our stakeholder in the product management department would like us to produce a short summary of sales in October. Their request includes a business rule: exclude any brands which sell less than $100 during the month.

This should filter out low-dollar items, such as the ACCESORY category, which will be handled in other reporting.

The Dangerous Query

Here's the first query we're going to try:

```
SELECT Brand, Sum(Qty) AS Total_Qty, Sum(PaidAmt) AS Total_Paid
FROM SalesTransactions (nolock)
WHERE PaidAmt >= 100
GROUP BY Brand
ORDER BY Brand
```

Not just bad, but dangerous. Why? Because it actually returns results that look valid, but they're quite incorrect:

	Brand	Total_Qty	Total_Paid
1	COATS	6	1364.70
2	GLOVES	6	312.90
3	PANTS	4	554.20
4	SCARVES	4	440.00

Since we're only talking about 30 records, let's copy the data to Excel and whip up a quick pivot:

Row Labels	Sum of PaidAmt
COATS	1,364.70
GLOVES	521.50
PANTS	554.20
SCARVES	440.00
SOCKS	160.93
ACCESORY	17.50
Grand Total	**3,058.83**

Our query came up with the right answer for Coats, Pants, and Scarves. It excluded Accessory, as intended, since our total in that brand was less than $100 for the time period. But the total for Gloves is incorrect, and the Socks brand is missing entirely.

What gives?

The problem is that the WHERE condition, PaidAmt >= 100, is applied at the record level before any aggregation occurs.

Looking back at our practice data, every record with a sale of Accessories or Socks is filtered out because no single invoice includes a PaidAmt of $100 or more.

Our sales of Gloves are a bit more of a problem. An invoice for a single pair goes for $52.15, while an invoice where two pair were bought has a paid amount of $104.30. Thus, single pair purchases are filtered out but purchases of two pair are included in the aggregation.

Since the Coats, Pants, and Scarves are correct, and the Gloves look like they *could* be correct, this is a dangerous query. Without scrutiny the information appears accurate and could lead to misguided decisions.

The Correct Query

Here's the correct query. The business rule of a minimum $100 threshold for the Brand is applied to the aggregated data, via HAVING.

```
SELECT Brand, Sum(Qty) AS Total_Qty, Sum(PaidAmt) AS Total_Paid
FROM SalesTransactions (nolock)
GROUP BY Brand HAVING Sum(PaidAmt) >= 100
ORDER BY Brand
```

And the results:

	Brand	Total_Qty	Total_Paid
1	COATS	6	1364.70
2	GLOVES	10	521.50
3	PANTS	4	554.20
4	SCARVES	4	440.00
5	SOCKS	7	160.93

In Conclusion

As the smartest man at Microsoft was fond of saying, "I'd rather have no data than bad data." Bad data leads to bad information, which leads to bad decisions.

Simple rule: WHERE applies before aggregating, HAVING applies after aggregating.

And no matter how confident you are in your coding skills, an understanding of the business is essential for reviewing a report and realizing that something simply doesn't look right. Significant errors don't always result in error messages.

Jeffrey Turner | LinkedIn (https://www.linkedin.com/in/jeffreyturnertx/)

