Data aggregation and interpretation metrics

Data aggregation is how an analyst **collects data** from **multiple sources** and **stores** it in an **understandable form**. For example, many rows of data in a spreadsheet can be summarized by the mean and variance of each row. Descriptive statistics are very helpful when performing data aggregation.

Data aggregation is another important part of the data exploration process. It helps analysts find trends in the data, make comparisons, and discover information that might have gotten lost in all the individual data points.

Data interpretation is the process of giving meaning to processed and analyzed data. It involves the following steps:

- Designing strong research questions
- Collecting data relevant to the questions you want to answer
- Analyzing collected data
- Summarizing the key findings of an analysis to answer research questions
- Reporting findings and conclusions

Data processing techniques, such as data filtering and searching, are essential in the process of data interpretation.

Data filtering involves splitting up the sample into groups to create new subsets to be analyzed.

Data searching helps to find specific records, e.g., unique values or specific strings, in a dataset.

Data is often aggregated using descriptive statistics, e.g., measures of frequency, central tendency, dispersion, and position.

In the following topics, we'll examine some of the common statistical measures used to aggregate data.

Count

The count or frequency of an item is the number of times the item occurs in a dataset.

Example

Find the number of dealerships that contain models of cars in each color in the following data:

Dealership	**Model**	**Color**	**Number in Stock**	**Miles Per Gallon (MPG)**
Velocity Motors	Corvette	Red	2	19
Elite Auto Group	Corvette	Red	2	19
Summit Motors	Model X	Red	3	102
Velocity Motors	GT-R	Blue	1	16
Precision Automotive	Civic	Blue	3	31
Elite Auto Group	Jetta	Green	2	29
Precision Automotive	Mustang	Green	2	21
Velocity Motors	Accord	Black	2	30

The data can be organized as in **Table 3-3**:

Color	Count

black	1
blue	2
green	2
red	3

The table gives the count or frequency of each car color in the data.

Implementation In R And Excel

	Code implementation

R	1. Store the data in a variable called x.
	2. Use the function `table()` to create a table of counts.
	
	In order to analyze the frequency or count of each unique value in a dataset, you can utilize the `table()` function in R. This function generates a frequency table, which provides a summary of how many times each distinct value occurs in the data.
	
	`x <- c("red", "red", "red", "red", "red", "red",`
	`"red", "blue", "blue", "blue", "blue", "green",`
	`"green", "green", "green", "black", "black")`
	`table(x)`
	
	This code produces the following output:
	
	`black blue green red`
	`2 4 4 7`

Excel

1. See the provided Excel spreadsheet for the data.
2. Click **Insert** > **PivotTable** > **From Table/Range**. The **PivotTable From Table/Range** dialog box appears.
   Select the range containing the data (i.e. `Sheet1!$A$1:$E$9`) and a cell in the Existing Worksheet in which to place the PivotTable(e.g., `$E$4`). Click **OK**.
3. Select **Color** as the row field and as the value field in the **PivotTable Fields** pane: The default function is Count.

Row Labels	Count of Color
Black	1
Blue	2
Green	2
Red	3
Grand Total	8

Sum

The sum of the number of red cars in stock at each location in a dataset is the result of adding the values in the Number In Stock feature for all cars of a specific color.

Example

Find the sum of all the red cars in stock at all locations in the dataset.

Implementation In R And Excel

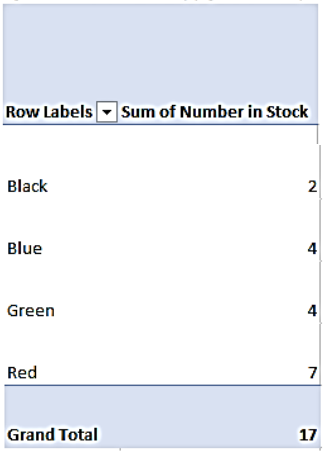
<table>
<tr><th></th><th>Code implementation</th></tr>
<tr><td>R</td><td>
1. Store the data in a variable called x.
2. Extract the count of red by using the function <code>table()</code>.

<pre>
x <- c("red", "red", "red", "red", "red", "red",
       "red", "blue", "blue", "blue", "blue", "green",
       "green", "green", "green", "black", "black")
color_table <- table(x)
red_count <- color_table["red"]
print(red_count)
</pre>
The code produces the following output:

<pre>
red
7
</pre>
</td></tr>
<tr><td>Excel</td><td>
1. Type the value <code>Red</code> in cell G1.
2. In cell H1, type the formula <code>=SUMIF(C:C, G1, D:D)</code>.
3. The formula locates each row that has <code>Red</code> in column C and sums the values in column D of those rows.
4. Type the value <code>Green</code> in cell G2 and fill the formula down to get the sum of dealerships that contain green cars.
5. Repeat Step 4 for <code>Blue</code> and <code>Black</code>.

OR

Use a pivot table:

1. See the provided Excel spreadsheet for the data.
2. Click Insert > PivotTable > From Table/Range.
The PivotTable From Table/Range dialog box appears.
Select the range containing the data (i.e. <code>Sheet1!$A$1:$E$9</code>) and a cell in the Existing Worksheet in which to place the PivotTable(e.g., <code>$H$4</code>).
Click OK.
3. Select <code>Color</code> as the row field and <code>Number In Stock</code> as the value field in the PivotTable Fields
4. By default, Excel will apply the <code>Sum</code> operation to the values. You should see the sum of the dataset displayed in the pivot table.


</td></tr>
</table>

Mean

The **mean** of a variable in a dataset is calculated by adding all of the values of the variable in the dataset and dividing their sum by the number of observations.

Video: How to calculate the mean

Example: The miles per gallon (MPG) of each of the car models for each dealership are below:

```
Velocity Motors: [19, 16, 30]

Elite Auto Group: [19, 29]

Summit Motors: [102]

Precision Automotive: [31, 21]
```

Find the average MPG for each dealership. The average MPG for `Velocity Motors` ($\bar{x}$) is below:

$$\text{Velocity Motors MPG}$$

$$\bar{x} = \frac{19 + 16 + 30}{3}$$

$$= \frac{65}{3} = 21.66666667\text{MPG}$$

Implementation In R, Excel, And SQL

	Code implementation

R	1. Store the data in a variable called `x`.
	2. Use the function `mean()` to find the mean.
	`x <- c(19, 16, 30)`
	`mean(x)`
	The code produces the following output:
	`[1] 21.66666667`

Excel	1. See the provided Excel spreadsheet for the data. 2. Use the `AVERAGEIF` function where `AVERAGEIF(range, criteria, [average_range])`. To get the average MPG for `Velocity Motors`, use the formula `=AVERAGEIF(A2:A9,"Velocity Motors",E2:E9)` 3. To get the average MPG for `Elite Auto Group`, use the formula `=AVERAGEIF(A2:A9,"Elite Auto Group",E2:E9)` 4. Repeat for the other dealerships. Tip: You can take a shortcut by typing the dealership names in a column and referencing that column in the formula instead of typing the name directly into the formula. OR Use a pivot table 1. See the provided Excel spreadsheet for the data. 2. Click **Insert** > **PivotTable** > **From Table/Range**. The **PivotTable From Table/Range** dialog box appears. Select the range containing the data (i.e. `$A$1:$E$9`) and a cell in the Existing Worksheet in which to place the PivotTable(e.g., `$M$4`). Click **OK**.

3. Select **Dealership** as the row field and **Miles Per Gallon (MPG)** as the value field in the **PivotTable Fields**

4. By default, Excel will apply the "Sum" operation to the values. Click on **Miles Per Gallon (MPG)** in the **PivotTable Fields** > **Value Field Settings** to change the operation to Average.

Row Labels	Average of Miles Per Gallon (MPG)
Elite Auto Group	24
Precision Automotive	26
Summit Motors	102
Velocity Motors	21.66666667
Grand Total	**33.375**

SQL

1. Open SQL Server Management Studio.
2. If you have not already created a database for running the examples in this course, create one.
3. Select the database and open the query window.

Note: The cars.sql script downloaded with the course files contains the statements in steps 4 and 5. You might find it easier to load that file than typing the statements.

1. Create a table called `CarDealerships`.

```
CREATE TABLE CarDealerships (
  Dealership VARCHAR(50),
  Model VARCHAR(50),
  Color VARCHAR(50),
  NumberInStock INT,
  MPG INT
);
```

2. Insert the values into the created table using the `INSERT INTO` statement.

```
INSERT INTO CarDealerships (Dealership, Model, Color, NumberInStock, MPG)
VALUES
    ('Velocity Motors', 'Corvette', 'Red', 2, 19),
    ('Elite Auto Group', 'Corvette', 'Red', 2, 19),
    ('Summit Motors', 'Model X', 'Red', 3, 102),
    ('Velocity Motors', 'GT-R', 'Blue', 1, 16),
    ('Precision Automotive', 'Civic', 'Blue', 3, 31),
    ('Elite Auto Group', 'Jetta', 'Green', 2, 29),
    ('Precision Automotive', 'Mustang', 'Green', 2, 21),
    ('Velocity Motors', 'Accord', 'Black', 2, 30);
```

3. Calculate the mean MPG of each dealership using the `GROUP BY` function.

```
SELECT Dealership, AVG(MPG) AS AverageMPG
FROM CarDealerships
GROUP BY Dealership;
```

4. The result will give you the average MPG for each dealership:

```
Elite Auto Group|24
Precision Automotive|26
Summit Motors|102
Velocity Motors| 21
```

Median

The **median** of a quantitative variable in a dataset is the value in the middle of the data when it is arranged in ascending order. The following steps can be used to find the median:

- Arrange the data in ascending order
- Determine the number (n) of observations
- If n is odd, the median is the middle observation. This observation can be found by dividing the number of observations by 2 and rounding it up. However, if n is even, the median is the mean of the two middle observations. Find these by dividing n by 2, and taking the average of the $\frac{n}{2}$ and $\frac{n}{2} + 1$ observations.

Example: Find the median MPG for all the car models on the lot at Velocity Motors.

Step 1: Arrange the data in ascending order, i.e., 16, 19, 30.

Step 2: The number of observations is 3.

Step 3: Because n is odd, the median is the middle number, which is 19.

The median MPG is 19 MPG.

Implementation In R And Excel

	Code implementation

R	1. Store the data in a variable called `x`.
	2. Use the function `median()` to find the median.
	
	`x <- c(16, 19, 30)`
	`median(x)`
	
	This code produces the following output:
	
	`[1] 19`

Excel	1. See the provided Excel spreadsheet for the data. Create the formula `=MEDIAN(E2,E5,E9)` in cell H1. 2. The formula outputs the required median in cell H1.

Mode

The **mode** of a variable is the most frequent observation of the variable in the dataset.

Example: Find the mode of MPG of all the car models at all dealerships.

You can summarize the data according to frequencies, as shown in **Table 3-4**. Use this table to determine the most frequently occurring data point.

Table 3-4

Item	Frequency
16	1
19	2
21	1
29	1
30	1
31	1
102	1

The mode of the MPG data is 19.

Implementation In Excel

	Code implementation
Excel	1. See the provided Excel spreadsheet for the data. 2. Click **Insert** > **PivotTable** > **From Table/Range**. The **PivotTable From Table/Range** dialog box appears. Select the range containing the data and a cell in the Existing Worksheet in which to place the PivotTable. Click **OK**. 3. Select **Miles Per Gallon (MPG)** as the row and value field in the **PivotTable Fields**. Change the value of MPG to be "Count" by selecting **value field settings** in the dropdown. The count gives you the number of occurrences of the value. 4. 19 MPG has the highest count of 2, so 19 MPG is the mode.

Range

The **range** of a variable is the difference between the largest and the smallest data value. To calculate the range, the variable must be quantitative, meaning that its values are numbers that describe an amount of something.

Example: Find the range of the MPG of all car models at all dealerships.

The largest MPG value (also called the maximum value) is 102.

The lowest MPG value (also called the minimum value) is 16.

The range of the MPG data is 102 - 16 = 86.

Implementation In R, Excel, And SQL

<table>
<tr><th></th><th>Code Implementation</th></tr>
<tr><td>R</td><td>
1. Store the data in a variable called x.
2. Use the function <code>max()</code> to find the maximum value of the data.
3. Use the function <code>min()</code> to find the minimum value of the data.
4. The range is the difference between the maximum and minimum values of the data.

<pre>x <- c(19, 19, 19, 19, 102, 102, 102, 16, 31, 31, 31, 29, 29, 21, 21, 30, 30)
rg <- max(x) - min(x)
rg</pre>
This code produces the following output:

<pre>[1] 86</pre>
</td></tr>
<tr><td>Excel</td><td>
1. See the provided Excel spreadsheet for the data.
2. Create the formula <code>=MAX(E2:E9)- MIN(E2:E9)</code> in cell H2.

This formula uses the functions <code>MAX</code> and <code>MIN</code> to calculate the maximum and the minimum values of the data, respectively. The range is the difference between these two values.

3. The formula outputs the required range in H2.

To find the minimum and maximum MPG for each dealership, you can use a pivot table.

1. See the provided Excel spreadsheet for the data.
2. Click Insert > PivotTable > From Table/Range.
The PivotTable From Table/Range dialog box appears.
Select the range containing the data and a cell in the Existing Worksheet in which to place the PivotTable.
Click OK.
3. Select Dealership as the row field and Miles Per Gallon (MPG) as the value field in the PivotTable Fields. Change the value of MPG to be Min by selecting value field settings in the dropdown.
4. Add Miles Per Gallon (MPG) again as a value field. Change the value field setting to Max. You will get the following result:
<table>
<tr><th>Row Labels</th><th>Min of Miles Per Gallon (MPG)</th><th>Max of Miles Per Gallon (MPG)</th></tr>
<tr><td>Elite Auto Group</td><td>19</td><td>29</td></tr>
<tr><td>Precision Automotive</td><td>21</td><td>31</td></tr>
<tr><td>Summit Motors</td><td>102</td><td>102</td></tr>
<tr><td>Velocity Motors</td><td>16</td><td>30</td></tr>
<tr><td>Grand Total</td><td>16</td><td>102</td></tr>
</table>
</td></tr>
<tr><td>SQL</td><td>
1. Launch SSMS.
2. Select the database you created and open a query window.
3. Execute the following query.
<pre>SELECT Dealership,
Min(MPG) AS 'Minimum MPG',
Max(MPG) as 'Maximum MPG'
FROM CarDealerships
GROUP BY Dealership;</pre>
</td></tr>
</table>

Sample Standard Deviation

The sample **standard deviation (s)** is the average amount of variation in a dataset. It describes how far each value lies from the mean. A low standard deviation means that data is clustered around the mean. However, a high standard deviation indicates that data is more spread out.

Given n observations $x_1, x_2, \ldots, x_n$, the standard deviation of the data can be calculated using the formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n - 1}},$$

where $\bar{x}$ is the mean value of all observations and *n* is the number of observations.

Example: Find the variance and standard deviation of the MPG for all the different cars at all the dealerships.

From the previous example, we calculated that the mean is 33.375 MPG.

The number of observations is 8.

You can use **Table 3-5** to obtain the squared deviations from the mean:

Table 3-5

x_i	$x_i - 33.375$	$(x_i - 33.375)^2$
19	-14.375	206.640625
19	-14.375	206.640625
102	68.625	4709.390625
16	-17.375	301.890625
31	-2.375	5.640625
29	-4.375	19.140625
21	-12.375	153.140625
30	-3.375	11.390625
	Total	5613.875

The sample variance is $s^2 = \frac{5613.875}{8-1} = \frac{5613.875}{7} = 801.98$ (to 2 decimal places)

The sample standard deviation is $s = \sqrt{\frac{5613.875}{7}} = 28.32$ (to 2 decimal places).

The variance of the data is 801.98MPG and the standard deviation is 28.32MPG.

Implementation In R And Excel

<table>
<tr><td></td><td colspan="2">Code Implementation</td></tr>
<tr><td></td><td>Variance</td><td>Standard deviation</td></tr>
<tr><td>R</td><td>1. Store the data in a variable called <code>x</code>.
2. Use the function <code>var()</code> to find the variance. You can use the command <code>round(var(x), 2)</code> to round off the variance to 2 decimal places.

<pre>x <- c(19, 19, 102, 16, 31, 29, 21, 30)
round(var(x), 2)</pre>This code produces the following output:

<pre>[1] 801.98</pre></td><td>Use the same steps as outlined in the Variance column of this table. Replace the <code>var()</code> function with <code>sd()</code> to calculate the standard deviation.

<pre>x <- c(19, 19, 102, 16, 31, 29, 21, 30)
round(sd(x), 2)</pre>This code produces the following output:

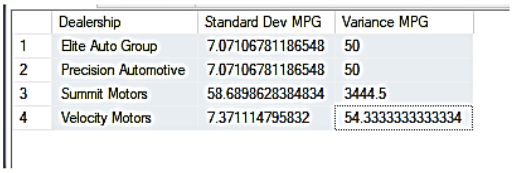
<pre>[1] 28.32</pre></td></tr>
<tr><td>Excel</td><td>1. See the provided Excel spreadsheet for the data.
2. Create the formula <code>=VAR(E2:E9)</code> in cell C10.
3. The formula outputs the required variance in cell C10.</td><td>1. See the provided Excel spreadsheet for the data.
2. Create the formula <code>=STDEV(E2:E9)</code> in cell D10.
3. The formula outputs the required standard deviation in cell D10.</td></tr>
<tr><td>Excel with a Pivot Table</td><td colspan="2">You can also use a pivot table to calculate variance and standard deviation.

1. Add the following row of data to your table.
Summit Motors Corvette Red 1 19
2. Create a pivot table from the data using the steps provided previously..
3. Add Dealership as the row value and Miles per Gallon (MPG) as the value field.
4. Change the value field settings to StdDev.
5. Add Miles per Gallon (MPG) as the value field again.
6. Change the value field settings to Var.

Your result should be as shown below.

<table>
<tr><th>Row Labels</th><th>StdDev of Miles Per Gallon (MPG)</th><th>Var of Miles Per Gallon (MPG)</th></tr>
<tr><td>Elite Auto Group</td><td>7.071067812</td><td>50</td></tr>
<tr><td>Precision Automotive</td><td>7.071067812</td><td>50</td></tr>
<tr><td>Summit Motors</td><td>58.68986284</td><td>3444.5</td></tr>
<tr><td>Velocity Motors</td><td>7.371114796</td><td>54.33333333</td></tr>
<tr><td>Grand Total</td><td>26.92014941</td><td>724.6944444</td></tr>
</table></td></tr>
</table>

	You can see that both standard deviation and variance for `Summit Motors` is much higher than for the other dealerships. This is expected because the range of values is 83. The two values are spread out; whereas the values for the other dealerships are clustered together.

SQL	1. Open SSMS.
	2. Select the database you created and open a query window.
	3. First, you will need to insert the new row for `Summit Motors`. This is important because if there is only one instance of a value in a dataset, the calculation for standard deviation and variance will require division by zero, which is an invalid operation.
	
	`INSERT INTO CarDealerships (Dealership, Model, Color, NumberInStock, MPG)`
	`VALUES (Summit Motors', 'Corvette', 'Red', 2, 19)`
	
	4. Execute the following query:
	
	`SELECT Dealership, StDev(MPG) AS 'Standard Dev MPG', Var(MPG) as 'Variance MPG'`
	`FROM CarDealerships`
	`GROUP BY Dealership;`
	
	5. Your results should be as follows:
	 

	Dealership	Standard Dev MPG	Variance MPG
1	Elite Auto Group	7.07106781186548	50
2	Precision Automotive	7.07106781186548	50
3	Summit Motors	58.6898628384834	3444.5
4	Velocity Motors	7.371114795832	54.3333333333334