

Data analysis (SQL)

Content

- Introduction to Data
 - Attributes of data (rows & columns)
 - Data collection
- Data sources
- Data types
- The data analyst's toolkit
- Database management systems
 - Types of DBMS
- Communicating with RDBMS

Content

- PostgreSQL
 - PostgreSQL installation (windows)
 - PostgreSQL installation (Mac OS)
 - Loading a database
- Writing SQL statements
 - Data definition language
 - Data manipulation language
 - Data control language
 - Transaction control language
 - Data querying language
- Primary and secondary keys

Content

- SQL syntax principles
- Retrieving data with the SELECT statement
 - Select, From and Where commands
- SQL aggregations
 - Count
 - Sum
 - Min & Max
 - Average
- Logical operators
- Special operators
- Temporary tables
- Case statements
- Date functions

Content

- Joins
 - Principles of Joins
 - Types of joins
 - Unions
- Order by and group by clauses
- SQL wildcards
- Subqueries
- String modification
- Data governance

Introduction to Data

So what exactly is data?

What is data?

“The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.” (Oxford dictionary, 2019).

Almost anything we can see is “data” to someone else.

Data is facts and statistics collected together for reference or analysis.

The root word for data is datum, a latin word meaning a piece of information or an assumption or premise from which inferences may be drawn.

What is a database?

A database is an organized collection of structured information, or data, typically stored electronically in a computer system.

A database is usually controlled by a database management system (DBMS).

Together, the data and the DBMS, along with the applications that are associated with them, are referred to as a database system, often shortened to just database.

Data within the most common types of databases in operation today is typically modeled in rows and columns in a series of tables to make processing and data querying efficient.

The data can then be easily accessed, managed, modified, updated, controlled, and organized. Most databases use structured query language (SQL) for writing and querying data.

Attributes of data (rows & columns)

A row of data is data stored horizontally while a column refers to data stored vertically.

Data within the most common types of databases in operation today is typically modeled in rows and columns in a series of tables to make processing efficient.

This way, data can then be easily accessed, managed, modified, updated, controlled, and organized. Most databases use structured query language (SQL) for writing and querying data.

Data Collection

Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.

“Data collection” is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.

The main goal of data collection is to collect information-rich data.

Question slide

Which of these are attributes of data?

1. It must contain numbers
2. It must have rows and columns
3. It may contain images, text and videos
4. All of the above

Data sources

Where do we get data from?

Data sources

Data sources are broadly divided into two types: Primary & Secondary data

Primary data is raw, original, and extracted directly from the sources.

Primary data is collected directly by performing techniques such as questionnaires, interviews, and surveys.

Secondary data is the data which has already been collected and reused again for analytic purpose.

Secondary data is previously recorded from primary data for example internal source and external source.

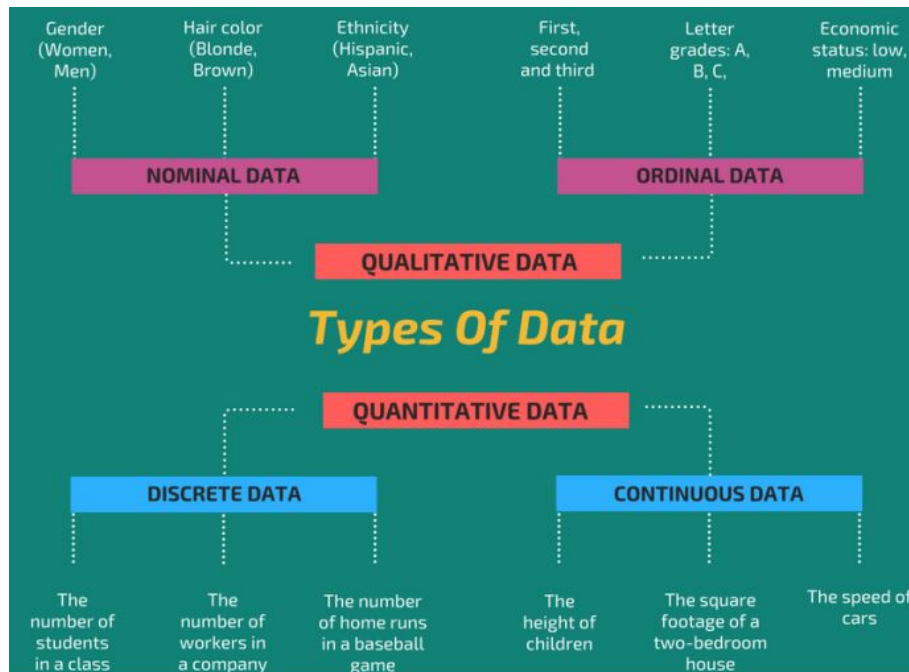
Data types

What are the types of data?

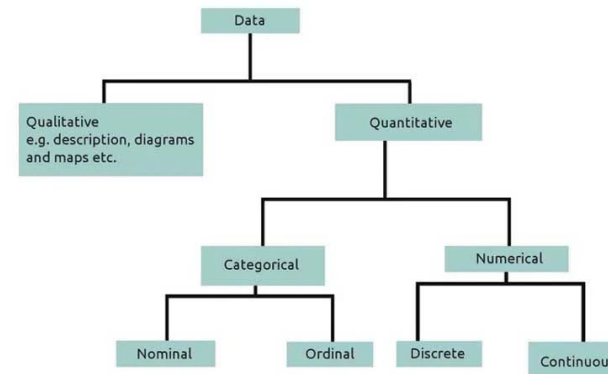
Qualitative & Quantitative data

Qualitative data is a group of non-numerical data such as words, sentences. It mostly focuses on behavior and actions of the group. Examples: names, sex

Quantitative data is in numerical forms and can be calculated using different scientific tools and sampling data. Examples: height, weight



Data Type	Type of Data	Memory Usage
Integer	An integer is a numeric variable without a decimal. Are whole numbers and can be positive, negative or zero, such as: 0, 2, 33, -199	2 or 4 bytes.
Real (Float)	Real numbers include all of the integer numbers that exist to infinity, plus all of their fractions and decimals. Such as: 1.26, -7.8, 3.14	4 or 8 bytes.
Char/Character	A character data type is used to store a single alphanumeric character. Where a character can be any letter, number or symbol that can be typed.	1 byte.
String	A string is more useful than the character data type as it can hold a list of characters of any length. Therefore it can represent alphanumeric data and symbols. A string can be null (empty), just one character or many characters.	1 byte per character in the string.
Boolean	A Boolean data type can only represent two values: True or False.	1 byte.



Question slide

Which of these are true?

1. There are three types of data sources
2. Numbers are always categorical data
3. Nominal data has a true zero point
4. Ordinal data has a true zero point

Data tools

What do we use to handle data?

SQL

Full meaning: Structured query language

History: SQL was first developed as “SEQUEL (Structured English query language)” in the 1970s by IBM researchers Raymond Boyce and Donald Chamberlin.

Sequel later became popularly known as SQL in the 1990s.

General uses: Querying databases.

Common types: NoSql, MySql, PostgreSQL...e.t.c.

Excel

Microsoft Excel is a spreadsheet application developed by Microsoft.

Uses: Calculation, graphing tools, pivot tables, macro programming via Visual Basic for Applications (VBA).

Power BI & Tableau

These are data visualization tools built by Microsoft and Tableau respectively.

They are mainly used for data visualization and business intelligence solutions.

They are very important for creating charts and graphs that can be directly linked to the database.

Python

Python is an interpreted high-level general-purpose programming language.

Its design philosophy emphasizes code readability with its use of significant indentation.

Python could be used in data science to query data, analyse data, create models and build machine learning algorithms.

Database management systems

How do we store data?

Database management systems

A database is an organized collection of data stored and accessed electronically from a computer system.

Think of your personal email.
You've got lots of mails, right? That's data.

That data is then stored by google along with over a billion other emails in a Database.

Databases are highly organised systems for storing data

The work of the data analyst begins at the database.

A database management system (DBMS) is a software package designed to define, manipulate, retrieve and manage data in a database.

A DBMS generally manipulates the data itself, the data format, field names, record structure and file structure.

It also defines rules to validate and manipulate this data. This means every DBMS has a set of pre-installed rules that allows it to store data properly so that it can be easily retrieved.

For example, the reason why the email field on a google form will only accept a valid email is because the DBMS has been set to only accept email for that field.

Types of database management systems

There are about seven common types of DBMS but for the purpose of this lecture, we will look at only two;

Relational database and NoSql

NoSQL databases are databases that do not use SQL as their primary data access language.

Graph database, network database, object database, and document databases are common NoSQL databases.

NoSQL database does not have predefined schemas.

NoSQL databases are the perfect candidate for rapidly changing development environments.

NoSQL allows developers to make changes on the fly without affecting applications.

Examples include: ArangoDB, MongoDB, CouchBase, Neo4j

*Schemas describes both the organization of data and the relationships between tables in a given database.

Types of database management systems

In a relational database management system (RDBMS), the relationship between data is relational and data is stored in tabular form of columns and rows.

Each column of a table represents an attribute and each row in a table represents a record.

Each field in a table represents a data value.

Some of the popular RDBMS are Oracle, SQL Server, MySQL and SQLite.

RDBMS are built on Primary and Foreign keys

Communicating with RDBMS

Structured Query Language (SQL) is the language used to communicate with (*pronounced; **Query*** 🗨️) RDBMS.

This includes inserting, updating, deleting, and searching tables.

Relational databases work on each table that has a key field that uniquely indicates each row. These key fields can be used to connect one table of data to another. All of this is done using SQL

Relational databases are the most popular and widely used databases and SQL is the most common **language** for querying databases.

*SQL is a **LANGUAGE!** Give yourself time to learn it. And like a language, it can only be learnt via repetition and practice.

PostgreSql

What tool can we use to write SQL?

PostgreSql installation (windows)

There are three steps to complete the PostgreSQL installation:

Download PostgreSQL installer for Windows

First, you need to go to [the download page](#) of PostgreSQL installers on the EnterpriseDB.

Second, click the download link as shown:

Version	Linux x86-64	Linux x86-32	Mac OS X	Windows x86-64	Windows x86-32
12.3	N/A	N/A	Download	Download	N/A
11.8	N/A	N/A	Download	Download	N/A
10.13	Download	Download	Download	Download	Download
9.6.18	Download	Download	Download	Download	Download
9.5.22	Download	Download	Download	Download	Download
9.4.26 (Not Supported)	Download	Download	Download	Download	Download
9.3.25 (Not Supported)	Download	Download	Download	Download	Download

Step 1. Double click on the installer file, an installation wizard will appear and guide you through multiple steps where you can choose different options that you would like to have in PostgreSQL.

Step 2. Click the Next button

Step 3. Specify installation folder, choose your own or keep the default folder suggested by PostgreSQL installer and click the Next button

Step 4. Select software components to install

The PostgreSQL Server to install the PostgreSQL database server,

pgAdmin 4 to install the PostgreSQL database GUI management tool,

Command Line Tools to install command-line tools such as psql, pg_restore, etc. These tools allow you to interact with the PostgreSQL database server using the command-line interface,

Stack Builder provides a GUI that allows you to download and install drivers that work with PostgreSQL.

*For this lesson, you don't need to install Stack Builder so feel free to uncheck it and click the Next button to select the data directory

Step 5. Select the database directory to store the data or accept the default folder. And click the Next button to go to the next step.

Step 6. Enter the password for the database superuser (postgres). This should be a password you can remember. After entering the password, you need to retype it to confirm and click the Next button.

Step 7. Enter a port number on which the PostgreSQL database server will listen. The default port of PostgreSQL is 5432. You should use that.

Step 8. Choose the default locale used by the PostgreSQL database. If you leave it as default locale, PostgreSQL will use the operating system locale and that is ideal. After that click the Next button.

Step 9. The setup wizard will show the summary information of PostgreSQL. You need to review it and click the Next button if everything is correct. Otherwise, you need to click the Back button to change the configuration accordingly. The installation may take a few minutes to complete.

Step 10. Click the Finish button to complete the PostgreSQL installation.

There are three steps to complete the PostgreSQL installation:

3. Verify the installation

Search for pgAdmin4 from your application list.

When it opens, it will ask for your password.

It should open up and be ready for use now.

Welcome!

PostgreSQL installation (Mac OS)

To download the PostgreSQL installer:

First, visit the [PostgreSQL installer download page](#).

Then, download the PostgreSQL for mac OS.

To install PostgreSQL on macOS, you follow these steps:

First, launch the setup wizard by double-click the installer file.

Second, select the directory where the PostgreSQL will be installed and click the Next button.

Third, select the components that you want to install, uncheck the Stack Builder, and click the Next button

The PostgreSQL Server to install the PostgreSQL database server,

pgAdmin 4 to install the PostgreSQL database GUI management tool,

Command Line Tools to install command-line tools such as psql, pg_restore, etc. These tools allow you to interact with the PostgreSQL database server using the command-line interface,

Stack Builder provides a GUI that allows you to download and install drivers that work with PostgreSQL.

Fourth, specify a directory where PostgreSQL stores the data and click the Next button

Fifth, enter the password for the postgres user account. You should note down this password for logging in to the PostgreSQL database server later. After that, click the Next button.

Sixth, specify the port number on which the PostgreSQL server will listen. By default, PostgreSQL uses port number 5432.

Seventh, select the locale used by PostgreSQL. By default, PostgreSQL uses the locale of the current operating system.

Eighth, review the installation information. If everything looks correct, click the Next button to begin the installation.

Ninth, click the Next button to start installing the PostgreSQL database server on your computer.

It will take few minutes to complete the installation. Click the Finish button once the installation is completed.

Loading a database

First, launch pgAdmin4 from Launchpad.

Second, enter the password for the postgres user.

Third, right-click the PostgreSQL 14 and select Create > Database.. to open a dialog for creating a new database.

Fourth, enter dvdrental as the database, postgres as the owner, and click the Save button to create the dvdrental database.

Sixth, download the sample database [from here](#)

Seventh, right-click the dvdrental database and select the Restore... menu item:
(If you get an error message "Please configure the PostgreSQL Binary Path in the Preferences dialog"; follow the steps here to correct it.)

Eighth, select the directory as the Format, the directory that contains sample database as the Filename, and postgres as the Role name, and click the Restore button.

It will take few seconds to restore the sample database. Once the restoration completes, you will see a notification.

It means that you have successfully created the sample database and restored it from the downloaded file.

Your database is now loaded and ready to use

Writing sql statements

Writing SQL statements

SQL statements refers to very clear and simple “sentences” that we use to query a database.

SQL statements are written in applications or Integrated Development Environments (IDE).

IDEs could be cloud based (e.g Google BigQuery) or hosted in a physical machine (eg. PostgreSql)

SQL statements are generally grouped into five

DDL – Data Definition Language

DQL – Data Query Language

DML – Data Manipulation Language

DCL – Data Control Language

TCL – Transaction control language

Data Definition Language (DDL)

DDL consists of the SQL commands that can be used to define the database schema.

It simply deals with descriptions of the database schema and is used to create and modify the structure of database objects in the database.

DDL includes a set of SQL commands used to create, modify, and delete database structures **but not data**.

*Think of it as the commands the bricklayers use to build a house (a database) not the furniture in the house (the data in the database)

DDL commands

CREATE: This command is used to create the database or its objects (like table, index, function, views, store procedure, and triggers).

Example: `CREATE TABLE table_name (column_name DATA TYPE, student_name TEXT, matric_no INTEGER);`

DROP: This command is used to delete objects from the database.

Example: `DROP TABLE table_name;`

ALTER: This is used to alter the structure of the database.

Example: `ALTER TABLE table_name`

`ADD CONSTRAINT student_pkey PRIMARY KEY (column_name);`

TRUNCATE: This is used to remove all records from a table, including all spaces allocated for the records are removed.

Example: `TRUNCATE TABLE table_name;`

DDL commands

COMMENT: This is used to add comments to the data dictionary. It is usually signified by "--" for single line comments, "/* */" for multi-line comments and in-line comments

```
-- single line comment  
-- another comment  
SELECT * FROM table_name;
```

```
/* multi line comment  
another comment */  
SELECT * FROM table_name;
```

```
SELECT * FROM /* table_name; */ (in-line comment)
```

RENAME: This is used to rename an object existing in the database.

Example: ALTER TABLE table_name
 RENAME TO new_table_name;

Data Manipulation Language (DML)

DMLs deal with the manipulation of data present in the database. They control access to data and to the database.

DML commands:

INSERT : It is used to insert data into a table.

Example: `INSERT INTO table_name (column1, column2, column3,..) VALUES (value1, value2, value3,..);`

UPDATE: It is used to update existing data within a table.

Example: `UPDATE table_name SET column1 = value1, column2 = value2,..`

DELETE : It is used to delete records from a database table.

Example: `DELETE FROM table_name WHERE some_condition;`

Data Control Language (DCL)

DCL includes commands which mainly deal with the rights, permissions, and other controls of the database system.

DCL commands:

GRANT: This command gives users access privileges to the database.

REVOKE: This command withdraws the user's access privileges given by using the GRANT command.

Transaction control language (TCL)

TCLs are used when writing Transaction SQL (TSQL) statements

Examples include:

COMMIT which commits a transaction.

ROLLBACK which rolls back a transaction in case of any error occurs.

SAVEPOINT which sets a savepoint within a transaction.

SET TRANSACTION which specifies characteristics for the transaction.

Primary and Foreign keys

Primary key	Foreign key
A column that is used to ensure data in the specific column is unique.	A column or group of columns in a relational database table that provides a link between data in two tables.
It uniquely identifies a record in the relational database table.	It refers to the field in a table which is the primary key of another table.
Only one primary key is allowed in a table.	More than one foreign key are allowed in a table.
It is a combination of UNIQUE values and does not allow for NULL values (empty cells).	It can contain duplicate values and a table in a relational database and can contain empty cells.
It cannot be deleted from the parent table.	It can be deleted from the table.

SQL syntax principles

SQL syntax refers to SQL statements that help us execute commands in SQL.

We use SQL syntax to build statements ranging from simple one line statements to complex blocks of statements.

All SQL statements follow SQL syntax principles and therefore must contain at least:

SELECT, a column name, FROM, a table name

Example: SELECT Surname from Staff_table

SELECT Column name FROM Table name

SQL syntax principles

Commands are not case sensitive

Column names are case sensitive. "Name" is not the same as "name"

Pay attention to indentation (to be explained later in the course)

Dates in SQL are arranged Year-Month-Day-Hour-Minute-Seconds. 2021:02:03 is the Third of February 2021.

Strings should be quoted in inverted commas "". Actor is not the same as "Actor". Note that these are also case sensitive.

SELECT, FROM, and WHERE commands

SELECT * FROM actor

SELECT FROM and WHERE are always present in every SQL query

SELECT points the server to the column(s) we want to extract data from

FROM points the server to the table we want to extract the data from

SELECT * FROM actor WHERE actor_id < 5

WHERE is a filter of sorts. It filters the output data.

Query Editor Query History

```
1 select*from actor
```

Data Output Explain Messages Notifications

	actor_id [PK] integer	first_name character varying (45)	last_name character varying (45)	last_update timestamp without time zone
1	1	Penelope	Guinness	2013-05-26 14:47:57.62
2	2	Nick	Wahlberg	2013-05-26 14:47:57.62
3	3	Ed	Chase	2013-05-26 14:47:57.62
4	4	Jennifer	Davis	2013-05-26 14:47:57.62
5	5	Johnny	Lollobrigida	2013-05-26 14:47:57.62
6	6	Bette	Nicholson	2013-05-26 14:47:57.62
7	7	Grace	Mostel	2013-05-26 14:47:57.62
8	8	Matthew	Johansson	2013-05-26 14:47:57.62
9	9	Joe	Swank	2013-05-26 14:47:57.62
10	10	Christian	Gable	2013-05-26 14:47:57.62

Query Editor Query History

```
1 select * from actor where actor_id < 5
```

Data Output Explain Messages Notifications

	actor_id [PK] integer	first_name character varying (45)	last_name character varying (45)	last_update timestamp without time zone
1	1	Penelope	Guinness	2013-05-26 14:47:57.62
2	2	Nick	Wahlberg	2013-05-26 14:47:57.62
3	3	Ed	Chase	2013-05-26 14:47:57.62
4	4	Jennifer	Davis	2013-05-26 14:47:57.62

Question slide

Which of these are true?

1. All SQL statements must contain a FROM clause
2. Most DDL start with a SELECT clause
3. Primary keys can have repeated values in a column
4. Secondary keys can have repeated values in a column

SQL Aggregations

SQL aggregations are functions that calculate on a set of values and return a single value.

They include COUNT, SUM, MIN, MAX and AVG

I. COUNT gives the count of attributes in a column.

Example: **Select COUNT (first_name) as First_name from actor** returns the number of first names in the first_name column of the actor table.

II. SUM adds up the value in integer columns.

Example: **Select SUM(amount) as Total_amount from payment** returns the total amount paid by customers in the payment table.

III. MIN & MAX return the minimum and maximum values of a column in the dataset.

Example: **Select MIN(amount) as min_amount from payment** returns the lowest amount paid by a customer.

IV. AVG computes the average value of a set of values from the dataset.

Example: **Select AVG(amount) as avg_amount from payment** will return the average amount paid by the customer.

Things to note (aggregations)

Aggregate functions always use parenthesis “()”

Columns must be aliased using the AS command.

Logical operators

Logical operators are SQL functions that return true or false values or combine two or more true or false values.

Logical operators include AND, OR and NOT

- I. AND compares two boolean expressions and returns true when both are true

Example: **SELECT * FROM actor WHERE actor_id <5 AND actor_id >2** returns rows where the actor_id is less than 5 but greater than 2

- I. OR compares two boolean expressions and returns true when both one of them is true

Example: **SELECT * FROM actor WHERE actor_id <5 OR actor_id >2** returns rows where the actor_id is less than 5 or greater than 2

- I. NOT takes a single boolean expression and changes its value from false to true or true to false

Example: **SELECT * FROM actor WHERE actor_id != 5** returns rows where the actor_id is not equal to 5 (!= means "not equal to")

Special operators

IN: The IN operator checks a value within a set of values separated by commas and retrieve the rows from the table which are matching

Example: **SELECT * FROM actor WHERE actor_id IN (2,3,4,5)** returns rows where the actor_id is 2,3,4 or 5.

BETWEEN: The SQL BETWEEN operator tests an expression against a range. The range consists of a beginning, followed by an AND keyword and an end expression

Example: **SELECT * FROM actor WHERE actor_id BETWEEN 2 AND 5** returns rows where the actor_id is less than 5 or greater than 2

Temporary tables

Temporary tables exist temporarily on the server. They get deleted once the last connection to the server is closed.

Temporary tables are very useful in scenarios when we have a large number of rows in a permanent database table and we have to frequently use some rows of this table.

We can select those specific rows which we need again and again from the permanent table into a temporary table and run queries on it more efficiently.

A temporary table can be created in two ways;

one creates the table first and then inserts values in it.

Second, creates it while selecting records from a permanent table.

Further details are beyond the scope of this course.

Case statements

SQL CASE statement is like an if-then-else statement. It goes through conditions and returns a value when the first condition is met.

Once a condition is true, it will stop reading and return the result. If no conditions are true, it returns the value in the ELSE clause.

If there is no ELSE part and no conditions are true, it returns NULL.

SYNTAX:

CASE

WHEN condition1 THEN result1

WHEN condition2 THEN result2

WHEN conditionN THEN resultN

ELSE result

END;

Example:

```
SELECT actor_id
CASE
  WHEN actor_id = 2 THEN "Actor 2"
  WHEN actor_id = 3 THEN "Actor 3"
  ELSE "Not actor 2 or 3"
END AS actor_name
FROM actor
```

Date functions

SQL date functions are functions that query and transform date, datetime and timestamp columns.

Common date functions include DATEDIFF, DATEADD, CURRENTDATE and other functions used to extract date parts and convert timezones.

```
Example: SELECT DATEDIFF ('1997-12-31 23:59:59','1997-12-30'),  
         SELECT DATE_ADD('1999-01-01', INTERVAL 1 HOUR),  
         SELECT CONVERT_TZ('2004-01-01 12:00:00','+00:00','+10:00')
```

<https://www.postgresql.org/docs/9.1/functions-datetime.html>

Question slide

Which of these are true?

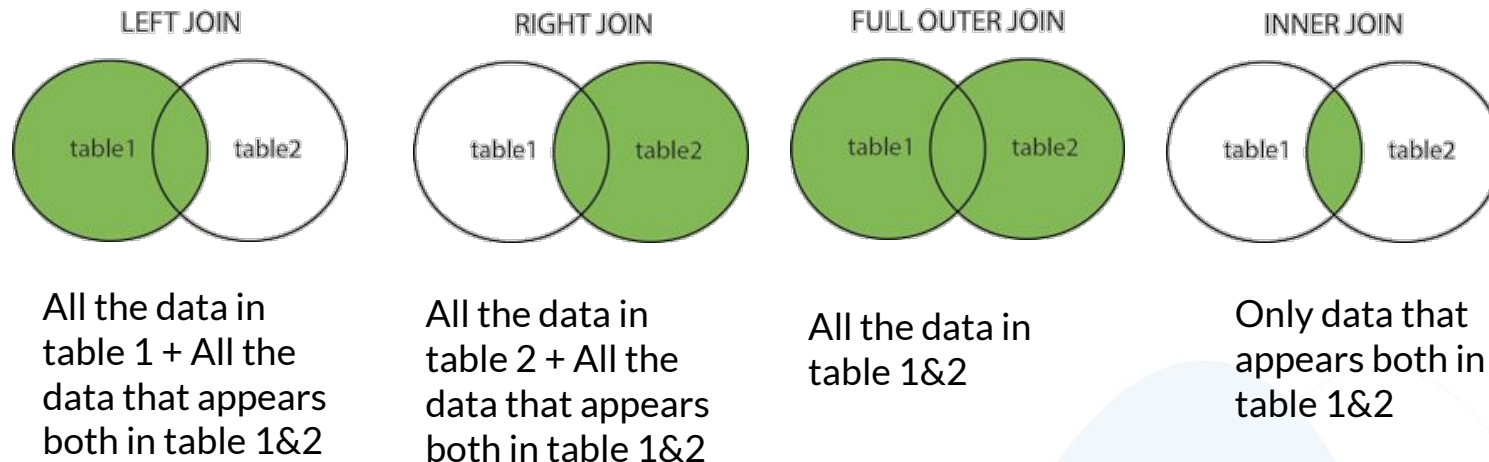
1. Every logical operator must use the GROUP BY clause
2. The IN and BETWEEN clause perform the same function
3. The WHERE and HAVING clause perform the same function
4. All of the above

Joins

Joins

Joins are used to query data from two or more tables.

There are different types of joins; Left join, Right join, Outer join and Inner join.



Joins

If Table 1 = actor table and table 2= film_actor table;

INNER JOIN:

```
SELECT actor.actor_id, actor.first_name,  
film_actor.film_id  
FROM actor  
INNER JOIN actor ON actor.actor_id =  
film_actor.actor_id;
```

RIGHT JOIN:

```
SELECT actor.actor_id, actor.first_name,  
film_actor.film_id  
FROM actor  
RIGHT JOIN actor ON actor.actor_id =  
film_actor.actor_id;
```

OUTER JOIN:

```
SELECT actor.actor_id,  
actor.first_name, film_actor.film_id  
FROM actor  
OUTER JOIN actor ON actor.actor_id =  
film_actor.actor_id;
```

LEFT JOIN:

```
SELECT actor.actor_id,  
actor.first_name, film_actor.film_id  
FROM actor  
LEFT JOIN actor ON actor.actor_id =  
film_actor.actor_id;
```

Principles of joins

Every column name in a join statement must be aliased using a dot notation (e.g. a.actor_id)

The type of join must be specified and the two columns of two different tables must be joined using the ON notation.

The two columns performing the join must be identical.

One or both of the columns in the join syntax must be identical.

Union

SQL union statements are used to join the results of two select statements together.

```
SELECT column_name(s) FROM table1  
UNION  
SELECT column_name(s) FROM table2;
```

The columns in both select statements must be the exact same.

```
SELECT column_name(s) FROM table1  
UNION ALL  
SELECT column_name(s) FROM table2;
```

The UNION operator selects only distinct values by default.
To allow duplicate values, use UNION ALL

Order by and Group by Clauses

The GROUP BY statement groups rows that have the same values into summary rows, like "find the number of customers in each country".

The ORDER BY statement orders the rows in our output data by ascending or descending order.

The GROUP BY and ORDER BY statements are often used with aggregate functions (COUNT(), MAX(), MIN(), SUM(), AVG()) to group or order the result-set by one or more columns.

```
SELECT * from customer GROUP BY store_id ORDER BY customer_id
```

SQL wildcards

A wildcard character is used to substitute one or more characters in a string.

Wildcard characters are used with the LIKE operator.

The LIKE operator is used in a WHERE clause to search for a specified pattern in a column.

LIKE Operator	Description
WHERE CustomerName LIKE 'a%'	Finds any values that starts with "a"
WHERE CustomerName LIKE '%a'	Finds any values that ends with "a"
WHERE CustomerName LIKE '%or%'	Finds any values that have "or" in any position
WHERE CustomerName LIKE '_r%'	Finds any values that have "r" in the second position
WHERE CustomerName LIKE 'a__%'	Finds any values that starts with "a" and are at least 3 characters in length
WHERE ContactName LIKE 'a%o'	Finds any values that starts with "a" and ends with "o"

Question slide

Which of these are true?

1. Joins can occur between more than two tables
2. Union queries can take different number of columns from its member tables
3. Every join must have an ON clause
4. SQL wildcards must be in the WHERE clause

Subqueries

Subqueries

A subquery or nested query is an SQL query that is nested into another SQL query

The subquery is also called an inner query or inner select, while the statement containing a subquery is also called an outer query or outer select

The subquery can be nested inside a SELECT, INSERT, UPDATE, or DELETE statement or inside another subquery.

A subquery is usually added within the WHERE Clause of another SQL SELECT statement. You can use the comparison operators, such as $>$, $<$, or $=$.

The inner query executes first before the outer query so that the results of an inner query can be passed to the outer query.

Subqueries

Example:

```
SELECT payment_id, customer_id, amount FROM payment  
WHERE amount > (SELECT avg(amount) FROM payment)
```

will return the payment table where the amount is higher than the average amount.

To create a subquery, first create your inner query then nest it into your outer query.

String modification

SQL has various functions for modifying string.

They include: Concat, Replace, Substr, Trim

Concat

The CONCAT SQL string function combines two or more strings into one string.

Syntax: *CONCAT(first_char, second_char, ... n_char)*

Example: *CONCAT("SQL", "is", "good")*

Output: SQL is good

Replace

SQL string. It returns an entry_char where the value of string_searching is replaced with string_replace.

Syntax: *REPLACE(entry_char, string_searching, string_replace)*

Example: *REPLACE("SQL is good", "good", "great")*

Output: SQL is great

String modification

SQL has various functions for modifying string.

They include: Concat, Replace, Substring, Trim

TRIM

The trim function in sql removes all the specific characters from a string.
Where no character is specified, the white spaces are removed.

Example: TRIM(" Cyber safe ")
Output: Cybersafe

SUBSTRING

Substring takes a part of a string and returns it

Syntax: *SUBSTR(char, position, length)*

Example: SUBSTR ("cybersafe", 6, 4)

Output: safe

Data governance

Data Governance and Profiling

Using SQL for Data Science



THANK YOU

Data analysis (Excel)

Content

- Introduction to Microsoft Excel
 - The Excel interface
 - Reading data into Excel
- Data preparation
 - Excel tables
 - Table formulas
 - Sorting
 - Filtering
 - Removing duplicates

Content

- Cell locking
- If Else Statement
- Vlookup
- Index match
- Excel syntax principles
- Conditional arithmetic procedures
- Central tendency
- Determining spread
 - Min and max
 - Quartiles

Content

- Running averages
- Histogram
- Scatter plots
- Forecasting
- Elementary data visualisation
- Principles of data visualization

Introduction to Microsoft Excel

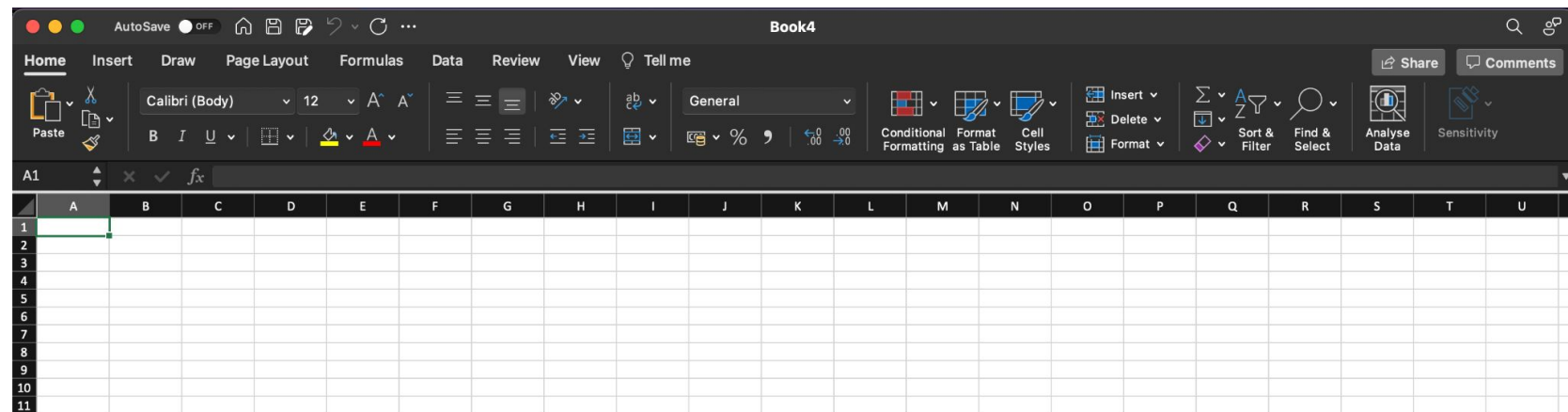
Introduction to Microsoft Excel

Microsoft Excel is a spreadsheet developed by Microsoft.

It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA).

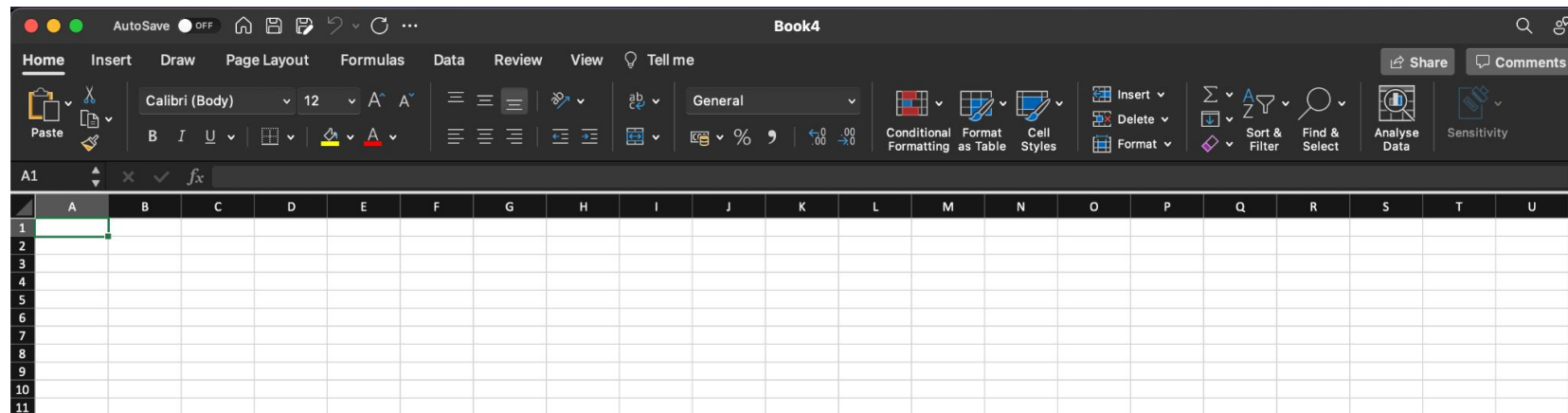
Since its release, it has become the industry standard for spreadsheets.

Excel forms part of the Microsoft Office suite of software.



The Excel interface

Row and column nomenclature

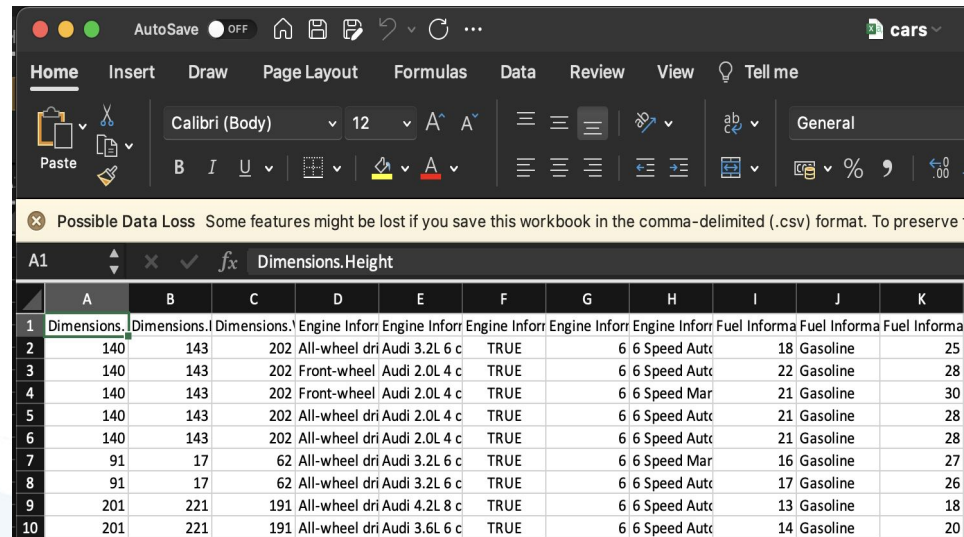


Reading data into Excel

Microsoft excel allows the user to read in data of various types into the interface. The most common types are Comma Separated Value (CSV), xlsx and text file.

Data is imported using the open tool in the file tab.

Once the data has been imported, it's rows and columns will be shown on the spreadsheet.



Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve

	A	B	C	D	E	F	G	H	I	J	K
1	Dimensions.	Dimensions.	Dimensions.	Engine Infor	Engine Infor	Engine Infor	Engine Infor	Engine Infor	Fuel Informa	Fuel Informa	Fuel Informa
2	140	143	202	All-wheel dri	Audi 3.2L 6 c	TRUE	6	6 Speed Autc	18	Gasoline	25
3	140	143	202	Front-wheel	Audi 2.0L 4 c	TRUE	6	6 Speed Autc	22	Gasoline	28
4	140	143	202	Front-wheel	Audi 2.0L 4 c	TRUE	6	6 Speed Mar	21	Gasoline	30
5	140	143	202	All-wheel dri	Audi 2.0L 4 c	TRUE	6	6 Speed Autc	21	Gasoline	28
6	140	143	202	All-wheel dri	Audi 2.0L 4 c	TRUE	6	6 Speed Autc	21	Gasoline	28
7	91	17	62	All-wheel dri	Audi 3.2L 6 c	TRUE	6	6 Speed Mar	16	Gasoline	27
8	91	17	62	All-wheel dri	Audi 3.2L 6 c	TRUE	6	6 Speed Autc	17	Gasoline	26
9	201	221	191	All-wheel dri	Audi 4.2L 8 c	TRUE	6	6 Speed Autc	13	Gasoline	18
10	201	221	191	All-wheel dri	Audi 3.6L 6 c	TRUE	6	6 Speed Autc	14	Gasoline	20

Data preparation

Data preparation

To optimise your use of various data management tools, you must make your data adhere to some basic standards. This is why we clean up and format data in excel.

Let us examine the PoorDesign worksheet in the data_prep file;

The title/headers are in multiple lines

There is an empty row

There is an hidden column...Can you find it?

There is inconsistent data formats in column F and H

Excel tables

Converting excel data to tables helps prevent problems with cleaning, sorting and formatting

Let us examine the Table Conversion worksheet in the data_prep file;

Before converting your data to table, you must...

1. Make sure the data is contiguous.
Select all the cells using control A
Use control . to move round the four corners of the data to check if there are empty cells or rows.
Do this step on the last worksheet we used. What do you notice?
2. Click on the Format as table icon in the styles group in the home tab.
Excel will automatically pick the range of your dataset and recognise the headers in it
Click ok
3. You'll notice the banded table style and the table design tab where you can change how the table looks

Table formulas

Converting excel data to tables helps prevent problems with cleaning, sorting and formatting

Let us examine the Table formulas worksheet in the data_prep file;

1. Create a new column New Comp. in column K. Notice the formatting
2. Calculate the new compensation which is the old compensation * the percentage increase.
Notice the formatting.
3. Press enter. Notice what happens

Calculate the average new compensation .

Let the output be in cell O3.

Use the table name in the formula.

What do you notice about the formula notation style?

Calculate the average new compensation .

Let the output be in cell O3.

Use the column range in the formula.

What do you notice about the formula notation style?

Use the method in the box 1 & 2 to the left to calculate the number of blank fields in column H using the countblank function

Sorting

You can sort data in excel using the sort button in the home tab or in the data tab.

The sort button in the data tab gives you more options.

We will be using the Sorting file for this lesson.

It's a great idea to make sure your list is in a table before sorting

To group the data by department, status and years of experience;

- Click the sort button in the data tab
- Add the columns you want to sort and the order
- Click ok

Filtering

Filtering helps you to see a selected section of your data

We will be using the Filtering file for this lesson.

With a filter, we can extract data that meet different criteria

Eg. People that work full time with a job rating of 5 in a particular department.

You can also filter by text. For example, filter for all the departments that contain the word “service”

Task:

- Filter for all the people hired in the first quarter of 2019
- Filter for all the people hired full-time between 2010 and 2015
- Filter for all the people that have compensation between 70000 and 90000
- Filter for the top ten compensation amounts

Removing duplicates

Sometimes, data comes with duplicate rows

The remove duplicate function helps us to remove these duplicate rows by specific rows.

We will use the Eliminate duplicate tab in the duplicates file for this lesson.

The remove duplicates button is under data tools in the data tab.

Use this button to remove the duplicate rows in the dataset.

You can also identify the duplicate rows by inserting a column into the dataset.

In that column, write a nested IF AND function to check if there are duplicates

=IF(AND(B2:L2=B3:L3), "dup", "unique")

Note that this is an array function. You may need to use ctrl+shift+enter to execute this command correctly.

Question slide

Which of these are true?

1. You can import images and videos into Excel
2. Microsoft Excel is the only application that can read spreadsheet data
3. All data must be turned into excel tables once imported
4. Excel tables cannot be formatted

Cell locking

Cell locking is the attribute of excel to pin a calculation or function to a specific cell during analysis.

We will look at the Cell locking file for examples on this lesson.

If Else statement

Excel If else statement returns an output based on multiple criterias

Using the IfElse file;

- I. Calculate the Total score for each student
- li. Create a column that returns
 - F for students that score below 29
 - E for students that score between 30 and 39
 - D for students that score between 40 and 49
 - C for students that score between 50 and 59
 - B for students that score between 60 and 69
 - A for students that score between 70 and above

Vlookup

Vlookup means Vertical lookup.

Vlookup is excel's method of looking into another table and getting a match of a desired value in the column.

I'll explain...with the Vlookup file.

Index & match

Index and Match serves as an alternative to vlookup.

Index and match are two formulas in excel that can be joined together to move data from one workbook to another.

We will use the Index-match file for this lesson.

Copy the Product column from the Product table to the Sales table

Copy the Category column from the Product table to the Sales table

Copy the Segment column from the Product table to the Sales table

Copy the ManufacturerID from the Manufacturer table column to the Sales table

Copy the Manufacturer Name from the Manufacturer table to the Sales table

Copy the State from the location table to the Sales table

Excel syntax principles

When writing syntax in excel, it is important to follow these steps;

1. Write the = sign
2. Start by typing the name of the function you want to perform (e.g; type SU and SUM will pop up)

Note that if there is no = sign, the function suggestion pane will not pop up.

Notice the difference in the two cells in the diagram

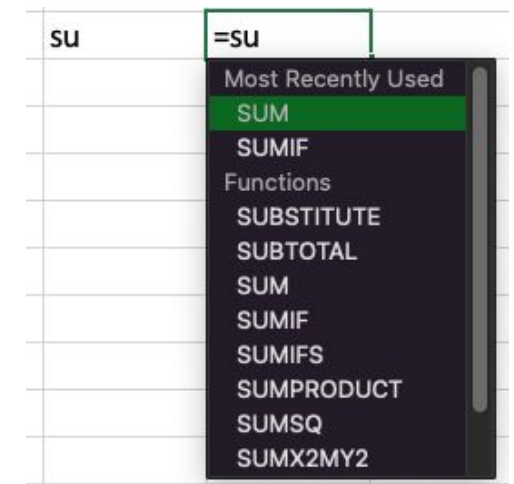
3. You do not have to type the function in full into your cell.

After typing = and the first few letters of your function, the function you want should be highlighted.

You can then press tab or enter to allow excel fill up the formula for you.

4. Your function must be followed by parenthesis “()”.

You will type what you want the function to do in the parenthesis/bracket.



Question slide

Which of these are true?

1. Hlookup and Vlookup perform the same function
2. Index and match are separate functions that can be joined together
3. Some excel formulas work without the = sign
4. None of the above

Conditional arithmetic procedures

Conditional arithmetic procedures

Conditional arithmetic procedures are arithmetic functions that use the IF clause
Most of the functions in excel can be found under the formula tab in excel
We will use the Con_arith file for this lesson.

Arithmetic functions include;

Countif: to count the number of full time workers in the dataset

Sumif: to calculate the total compensation of full time workers in the dataset

Averageif: to calculate the average compensation of full time workers in the dataset

Countifs: to count the number of full time workers with a job rating of 5 and have been in the company more than 10 years.

Averageifs: to calculate the average compensation of full time workers in the dataset with a job rating of 5 and have been in the company more than 10 years.

Maxifs: to calculate the Maximum compensation of full time workers in the dataset with a job rating of 5 and have been in the company more than 10 years.

Minifs: to calculate the minimum compensation of full time workers in the dataset with a job rating of 5 and have been in the company more than 10 years.

Central tendency

The most common operations in excel are the mean, median and the mode.

We will be using the MeansAndMedian file for this section.

In Cell D2, type the syntax, =AVERAGE(A2:A12). You should get an average of 29.4545

In Cell D4, type the syntax, =MEDIAN(A2:A12). You should get a median of 28

In Cell D6, type the syntax, =MODE(A2:A12). You should get a mode of 28

The mean, median and mode are measure of central tendency and they give us an idea of the "center" of the data.

They are affected by the skewness of data to varying degree.

Discussion: What do the other two types of modes do?

Activity: Change cell A11 to 31. Notice any changes in the output cells? Why do you think they changed?

Determining spread (Min & Max)

The most common operations in excel are the mean, median and the mode.

We will be using the MinMax file for this section

Remember to hold down the cntrl, shift key and down arrow key from cell A2 to select up to cell A41 .

In Cell D2, type the syntax, =MIN(A2:A41). You should get the minimum order value of 1684.00

In Cell D4, type the syntax, =MAX(A2:A41). You should get the minimum order value of 9932.00

Determining spread (Quartiles)

Quartiles divide your dataset into four segments;

The lowest quartile, Q1

The second quartile, Q2

The median

The fourth quartile, Q3

There are two ways to analyse quartiles in excel. The inclusive and exclusive method

The inclusive method:

In Cell D7, type the syntax, =QUARTILE.INC(A2:A41,1). 1 indicates the first quartile

The exclusive method:

In Cell G7, type the syntax, =QUARTILE.EXC(A2:A41,1). 1 indicates the first quartile

Classwork:

The results in each method is different. Why?

Complete the calculations for the other quartiles

Running averages

Running averages look at the average of data over a progressive time range

We will be using the RunningAverage file for this section

A.

In cell C2, use the average function to calculate the average in cell B2 alone. Use the absolute reference to make the average function always begin at cell B2.

Double click the fill handle to populate the other cells in column C.

How is this result calculated?

B.

In cell D4, use the average function to calculate the average from January to March

Double click the fill handle to populate the other cells in column D.

How is this result calculated?

Forecasting

Forecasting answers the question of what happens next based on data gathered on current trends

We will be using the forecast file for this section

Start with highlighting cells B2:B9 in the Trend worksheet

Drag the fill handle (the little green box at the tip of the last highlighted cell) down to row 13

Open the Forecast worksheet

Notice that the amount spent does not have a progression like the Trend sheet

To predict the amount a customer will spend if he travels 30 miles:

i. type 30 in cell D2

ii. in cell E2, use the FORECAST.LINEAR function.

Where;

X = miles driven,

Known_ys = Dependent variable = Amount spent

Known_xs = independent variable = Distance driven

Histograms

A histogram is a chart that shows the number of values in given ranges. Each range is called a bin. All bins have equal widths.

We will be using the Histogram file for this section

Procedure:

Click on a cell containing the values

Go to the **charts** group in the **insert** tab, click **histogram**.

You can right click on the chart and select the **format** option to change the format of the chart.

Change the width of the bins.

Scatter plots

A scatter plots gives the relationship between two values.

We will be using the XYScatter file for this section

Procedure:

Click on a cell containing the values

Go to the charts group in the insert tab, click scatter.

This shows a chart of the correlation between distance traveled to a store and the amount spent at the store.

You can right click on the chart and select the format option to change the format of the chart.

Note that correlation is not causation.

Scatter plots visualise numerical data types only

Task: Add a trendline to the chart. What does it mean?

Question slide

Which of these are true?

1. Running and normal averages produce different results
2. Running and normal averages produce the same results
3. The size of bins determine the number of bars in a histogram
4. All of the above

Elementary data visualisation

What is a dashboard?

- A dashboard is a visual display of the **most important information** needed to achieve one or more objectives; **consolidated** and arranged on a single screen so the information can be **monitored at a glance.** – Stephen few (Information dashboard design)

Functions of dashboards

- They optimize our understanding of complex systems and business processes through a consolidated format.
- To perform descriptive analysis
- To perform performance analysis in respect to time

Core Principles of Data Visualization

Audience



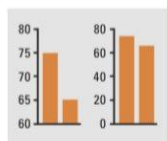
Always consider your audience—whether they need a short, written report, a more in-depth paper, or an online exploratory data tool.

Use pie charts with care



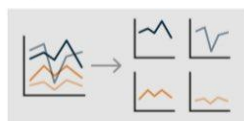
We are not very good at discerning quantities from the slices of the pie chart. Other chart types—for example, bars, stacked bars, treemaps, or slope charts—may be a better choice.

Start bar and column charts at zero



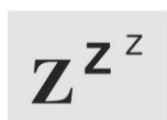
Bar and column charts that do not start at zero overemphasize the differences between the values. For small changes in quantities, consider visualizing the difference or the change in the values.

Try small multiples

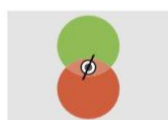


Breaking up a complicated chart into smaller chunks can be an effective way to visualize your data.

Color and font considerations



Avoid default colors and fonts—they all look the same and don't stand out.



Consider color blindness—about 10% of people (mostly men) have some form of color blindness.



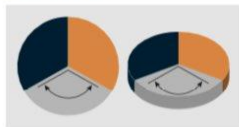
Avoid the rainbow color palette—it doesn't map to our number system and there is no logical ordering.

Include annotation



Add explanatory text to help the reader understand how to read or use the visualization (if necessary) and also to guide them through the content.

Avoid 3D



Using 3D when you don't have a third variable will usually distort the perception of the data and should thus be avoided.

Make labels easy to read



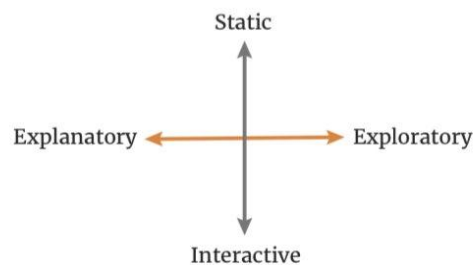
When applicable, rotate bar and column charts to make the labels horizontal. If possible, make vertical axis labels horizontal, possibly below the title. In general, make labels clear, concise, and easy for your reader to understand.

Use maps carefully



Use maps carefully, always being sure it is the geographic point you are trying to make. Column and bar charts, for example, are often better at enabling comparisons between geographic units.

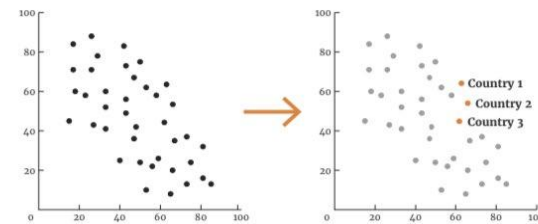
Visualization Mapping: Form and Function



Core Principles of Data Visualization

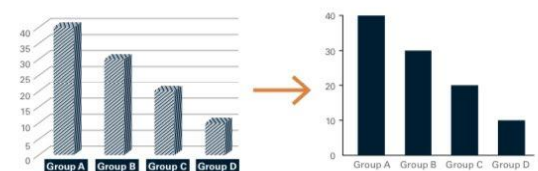
Show the data

People read graphs in a research report, article, or blog to understand the story being told. The data is the most important part of the graph and should be presented in the clearest way possible. But that does not mean that all of the data must be shown—indeed, many graphs show too much.



Reduce the clutter

Chart clutter, those unnecessary or distracting visual elements, will tend to reduce effectiveness. Clutter comes in the form of dark or heavy gridlines; unnecessary tick marks, labels, or text; unnecessary icons or pictures; ornamental shading and gradients; and unnecessary dimensions. Too often graphs use textured or filled gradients.

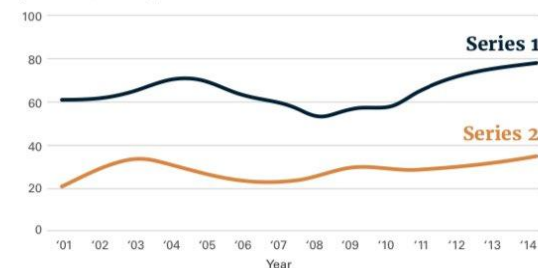


Integrate the text and the graph

Standard research reports often suffer from the **slideshow effect**, in which the writer narrates the text elements that appear in the graph. A better model is one in which visualizations are constructed to complement the text and at the same time to contain enough information to stand alone. As a simple example, legends that define or explain a line, bar, or point are often placed far from the content of the graph—off to the right or below the graph. Integrated legends—right below the title, directly on the chart, or at the end of a line—are more accessible.

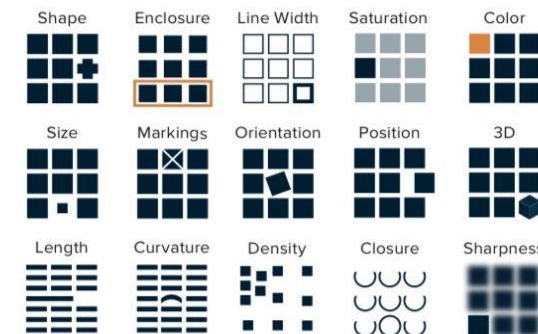
Chart Title Here

(Y axis label here)



Preattentive Processing

Effective data visualization taps into the brain's **preattentive visual processing**. Because our eyes detect a limited set of visual characteristics (such as shape and contrast), we combine various characteristics of an object and unconsciously perceive them as comprising an image. Preattentive processing refers to the cognitive operations that can be performed prior to focusing attention on any particular region of an image. In other words, it's the stuff you notice right away.



Steps in creating a visual

- Create an excel table
- Create a pivot table
- Visualise with pivot charts
- Slicers and timers

What are the principles of visualisation?
Types of charts and how to create them

We will be looking at the questions in the visualisation file, answering each question by creating pivot tables and charts with slicers.

Question slide

When creating a visual...

1. The design should be tailored to the analyst's taste
2. The design should be tailored to the end user's taste
3. Legends should be hidden
4. All of the above



THANK YOU