



DATA ENGINEERING

Module 2

Outline

- Data Engineering with Excel
- Data Engineering Review
- Excel vs SQL vs Python



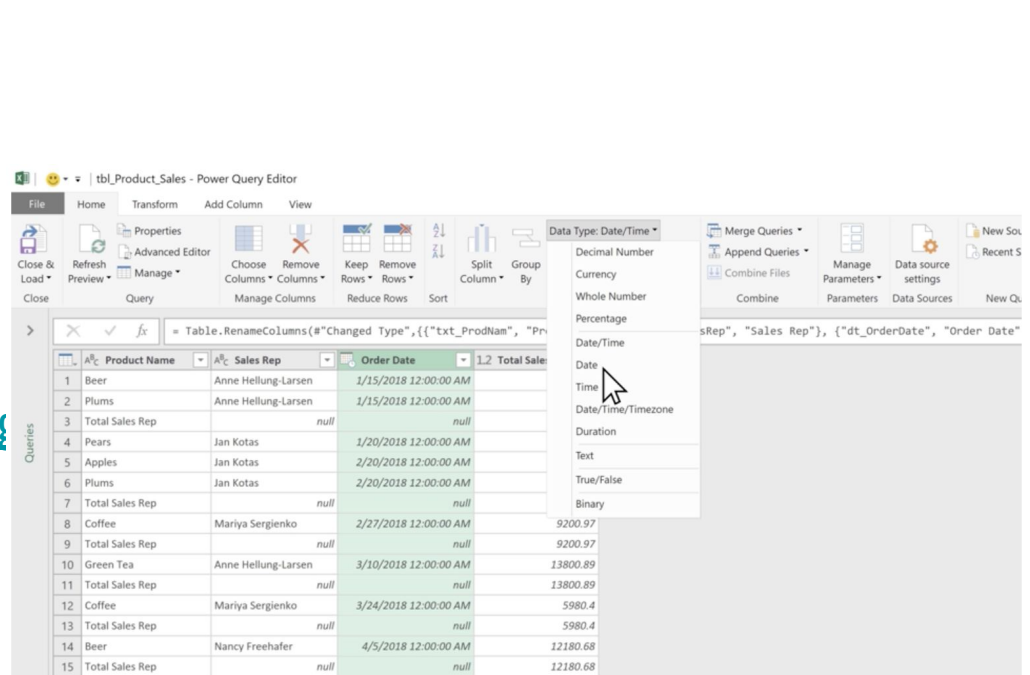
Data Wrangling Exercises with Pandas



Convert data type

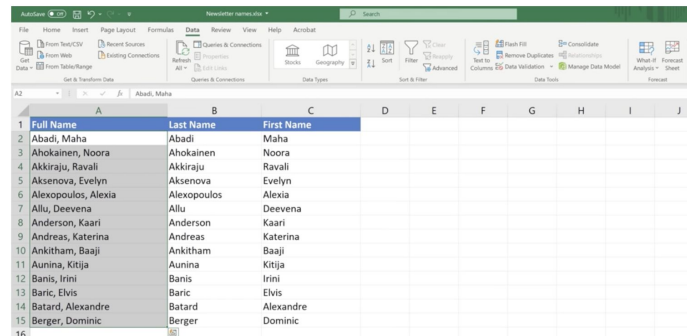
- Convert to data type

- Modify or change the data type settings of a field



Splitting text into columns

- Split text into columns



The screenshot shows an Excel spreadsheet with a table of names. The table has three columns: 'Full Name', 'Last Name', and 'First Name'. The data is as follows:

Full Name	Last Name	First Name
Abadi, Maha	Abadi	Maha
Ahokainen, Noora	Ahokainen	Noora
Akkinraju, Ravali	Akkinraju	Ravali
Aksenova, Evelyn	Aksenova	Evelyn
Alexopoulos, Alexia	Alexopoulos	Alexia
Allu, Deevana	Allu	Deevana
Anderson, Kaari	Anderson	Kaari
Andreas, Katerina	Andreas	Katerina
Ankitham, Baaji	Ankitham	Baaji
Aunina, Kitija	Aunina	Kitija
Baric, Irini	Baric	Irini
Baric, Elvis	Baric	Elvis
Batard, Alexandre	Batard	Alexandre
Berger, Dominic	Berger	Dominic

- Split text into columns with functions

	Example name	Description	First name	Middle name	Last name	Suffix
1	Jeff Smith	No middle name	Jeff		Smith	
2	Eric S. Kurjan	One middle initial	Eric	S.	Kurjan	
3	Janaina B. G. Bueno	Two middle initials	Janaina	B. G.	Bueno	
4	Kahn, Wendy Beth	Last name first, with comma	Wendy	Beth	Kahn	
5	Mary Kay D. Andersen	Two-part first name	Mary Kay	D.	Andersen	

Cleaning data using Excel

● Top ten ways to clean your data

Misspelled words, stubborn trailing spaces, unwanted prefixes, improper cases, and nonprinting characters make a bad first impression. And that is not even a complete list of ways your data can get dirty. Roll up your sleeves. It is time for some major spring-cleaning of your worksheets with Microsoft Excel.

The basics of cleaning your data	▼
Spell checking	▼
Removing duplicate rows	▼
Finding and replacing text	▼
Changing the case of text	▼
Removing spaces and nonprinting characters from text	▼
Fixing numbers and number signs	▼
Fixing dates and times	▼
Merging and splitting columns	▼
Transforming and rearranging columns and rows	▼
Reconciling table data by joining or matching	▼
Third-party providers	▼

Using Analyze Data in Excel

● Get insights with Analyze Data

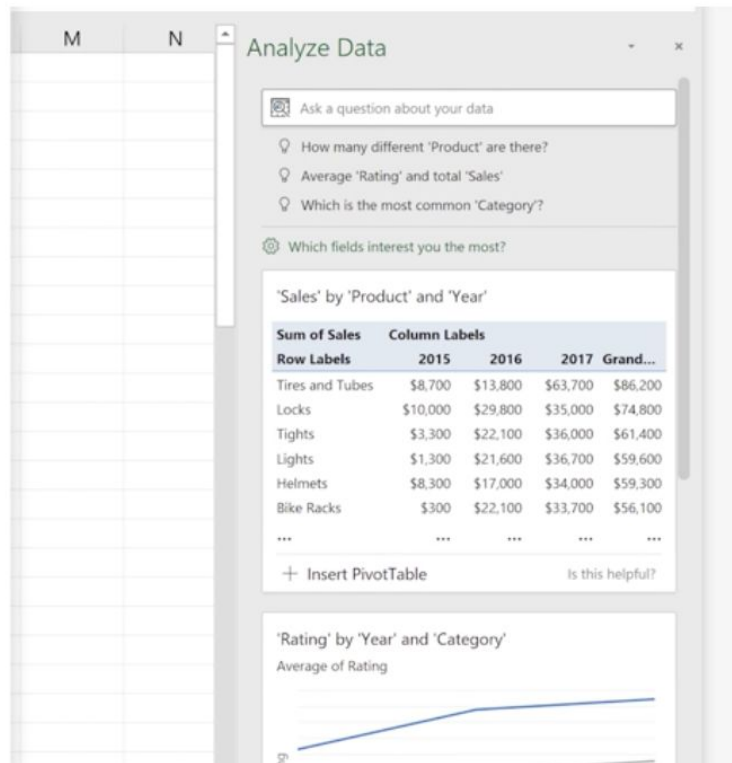
Analyze Data in Excel for the web helps you gain insights into your data through high-level visual summaries, trends, and patterns.

1. Select a cell in a data range.
2. Select **Home** > **Analyze Data**.

The **Analyze Data** pane will appear and show different visual and analysis types, such as:

- **Rank**
- **Trend**
- **Outlier**
- **Majority**

3. Choose an option and select **Insert PivotChart**.

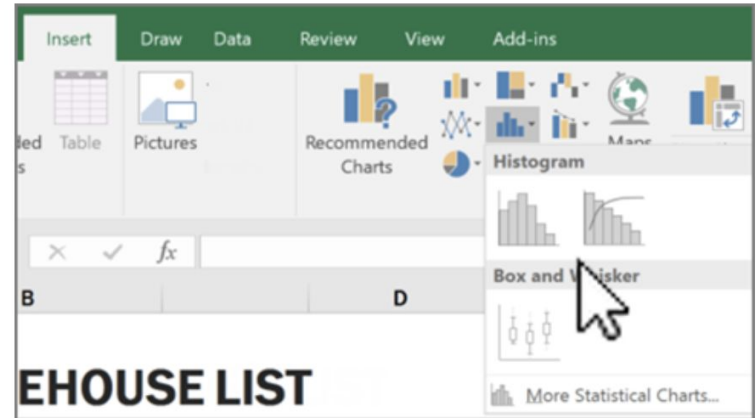


Histogram

- Creating histogram with Excel

Excel 2016

1. Select your data.
2. On the **Insert** tab, click **Insert Statistical Chart > Histogram**.



Statistical Functions

- Statistical Functions in Excel

Mean, Median, Mode
Variance, Standard Deviation, Skew, Kurtosis

Correlation, Covariance, Pearson

Function	Description
AVEDEV function	Returns the average of the absolute deviations of data points from their mean
AVERAGE function	Returns the average of its arguments
AVERAGEA function	Returns the average of its arguments, including numbers, text, and logical values
AVERAGEIF function	Returns the average (arithmetic mean) of all the cells in a range that meet a given criteria
AVERAGEIFS function	Returns the average (arithmetic mean) of all cells that meet multiple criteria

Loading the Analysis ToolPak in Excel


● in MacOS

● in Windows

1. Click the **Tools** menu, and then click **Excel Add-ins**.
2. In the **Add-Ins available** box, select the **Analysis ToolPak** check box, and then click **OK**.
 - a. If **Analysis ToolPak** is not listed in the **Add-Ins available** box, click **Browse** to locate it.
 - b. If you get a prompt that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.
 - c. Quit and restart Excel.

Now the **Data Analysis** command is available on the **Data** tab.

1. Click the **File** tab, click **Options**, and then click the **Add-Ins** category.

If you're using Excel 2007, click the **Microsoft Office Button** , and then click **Excel Options**
2. In the **Manage** box, select **Excel Add-ins** and then click **Go**.

If you're using Excel for Mac, in the file menu go to **Tools > Excel Add-ins**.
3. In the **Add-Ins** box, check the **Analysis ToolPak** check box, and then click **OK**.
 - If **Analysis ToolPak** is not listed in the **Add-Ins available** box, click **Browse** to locate it.
 - If you are prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.

Using Analytics ToolPak in Excel

- Use the Analysis ToolPak to perform complex data analysis

*The Analysis ToolPak includes the tools described in the following sections. To access these tools, click **Data Analysis** in the **Analysis** group on the **Data** tab. If the **Data Analysis** command is not available, you need to load the Analysis ToolPak add-in program.*

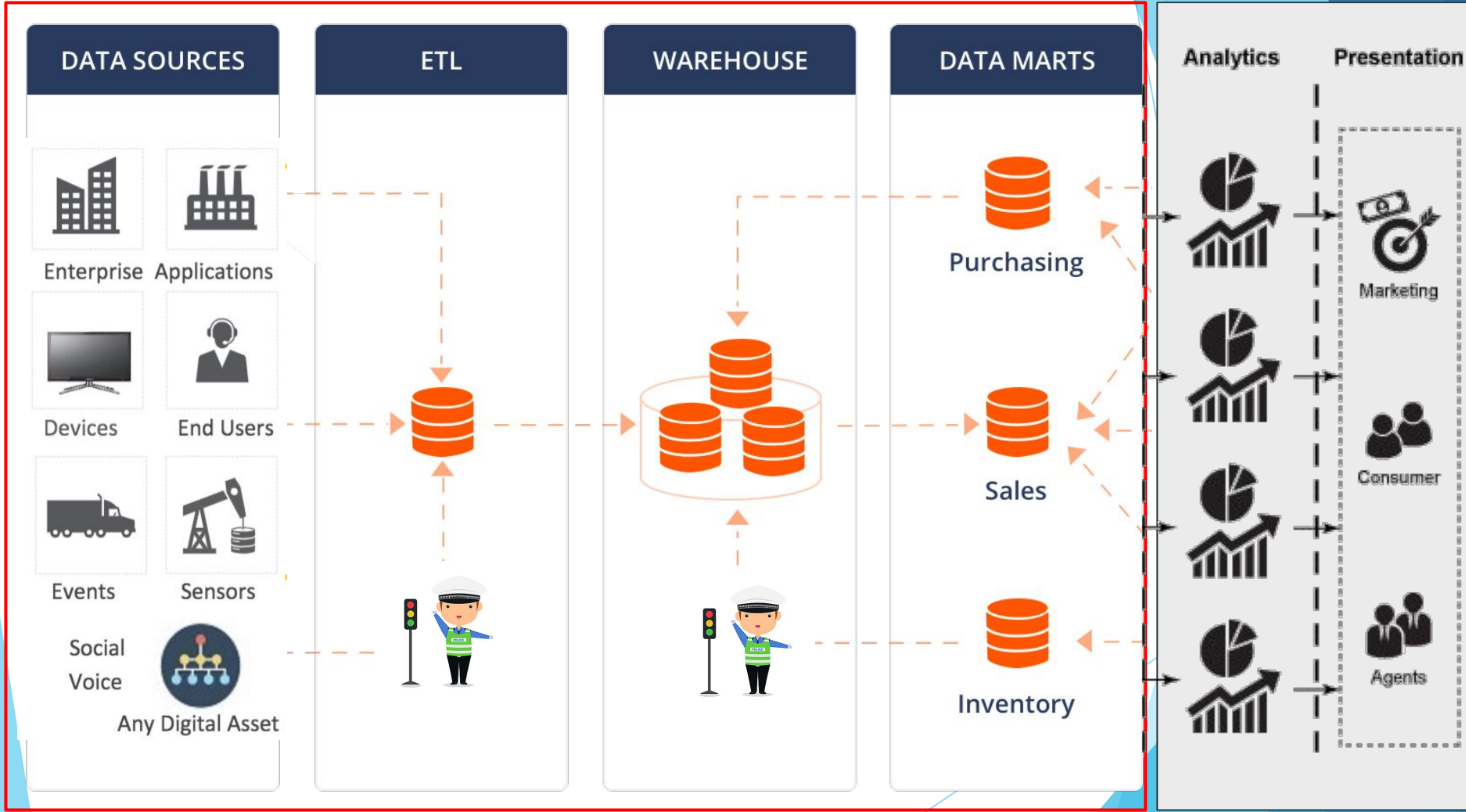
Load and activate the Analysis ToolPak	▼
Anova	▼
Correlation	▼
Covariance	▼
Descriptive Statistics	▼
Exponential Smoothing	▼
F-Test Two-Sample for Variances	▼
Fourier Analysis	▼
Histogram	▼
Moving Average	▼
Random Number Generation	▼
Rank and Percentile	▼
Regression	▼
Sampling	▼
t-Test	▼
z-Test	▼

movie dataset

1. apply cleaning techniques discussed
2. answer questions on the dataset
 - *What is the oldest film? Top gross? Top budget? Top net? longest film? longest series?*
 - *Who is the top grossing director? How many films?*
 - *Which genres are the top grossing films are?*
 - *Genre with the least count? Which films, and how much did they earn?*
 - *Are there any popularly casted actors in the films listed? If yes, who?*
 - *Which films were highly rated? Were they popular? Of high budget?*

Summary: Excel vs SQL vs Python







Excel



Python



MySQL

type

spreadsheet

programming
language

RDBMS

cost

licensed

open source

open source

data source

flat files; reports

API; web service

database; data
dump; logs

output

report; dashboard

flat file; database;
API

updated data;
dashboard



Excel



Python



MySQL

objective

simple calculations
& visualization

data analytics

data storage and
manipulation

analytics

descriptive

descriptive;
predictive

descriptive

streamline

lookups & macros

frameworks &
libraries

procedures &
triggers

scaling

software;
worksheets

CPU; memory;
libraries

disk space;
optimization

versioning

multiple copies

git

multiple tables



DATA ENGINEERING

Module 2