DATA AND
ARTIFICIAL INTELLIGENCE

# Data Science Capstone



simplilearn

DATA AND
ARTIFICIAL INTELLIGENCE

# Share Market Data Analysis



simplilearn

# Business Scenario

Problem statement:

The Nasdaq-100 is a stock market index comprised of 102 equity securities issued by 101 of the Nasdaq's largest nonfinancial companies. It includes sectors such as manufacturing, technology, retail, telecommunication, biotechnology, health care, transportation, media, and service providers.

The cluster trading strategy is used to build a diverse portfolio of investments. This method enables the identification of different company segments. One advantage of this analysis is that it can help to protect an investor's portfolio from risks.

Objective:

You must now create such segments so that customers can identify segments to invest in and segments to avoid. Use cluster analysis techniques to accomplish this task. You will also need to perform time-series forecasting for stock prices.

# Dataset Snapshot

## Nasdaq 100 Market cap.xlsx

| | Symb | Name | Market Cap | Last Sale | Net Change | Percentage Chan |
|---|---|---|---|---|---|---|
| 2 | AAPL | Apple Inc. | 26,25,74,01,43,000 | $151.45 | $2.00 | 1.34% |
| 3 | ABNB | Airbnb, Inc. | 69,56,99,44,167 | $116.65 | $0.26 | 0.22% |
| 4 | ADBE | Adobe Inc. | 1,49,14,45,69,000 | $320.81 | $4.59 | 1.45% |
| 5 | ADI | Analog Devices, Inc. | 75,48,47,63,090 | $146.76 | $2.23 | 1.54% |
| 6 | ADP | Automatic Data Processing, Inc. | 98,33,27,62,096 | $236.78 | $0.13 | -0.05% |
| 7 | ADSK | Autodesk, Inc. | 46,28,12,59,186 | $214.405 | $7.315 | 3.53% |
| 8 | AEP | American Electric Power Company, Inc. | 44,80,78,78,084 | $87.22 | $1.22 | 1.42% |
| 9 | ALGN | Align Technology, Inc. | 17,17,36,46,636 | $219.87 | $8.48 | 4.01% |
| 10 | AMAT | Applied Materials, Inc. | 75,73,73,36,004 | $88.035 | $3.095 | 3.64% |
| 11 | AMD | Advanced Micro Devices, Inc. | 99,90,21,45,714 | $61.3889 | $2.6889 | 4.58% |
| 12 | AMGN | Amgen Inc. | 1,39,43,50,75,361 | $260.66 | $0.66 | -0.25% |
| 13 | AMZN | Amazon.com, Inc. | 12,28,24,66,01,480 | $120.6898 | $0.8698 | 0.73% |
| 14 | ANSS | ANSYS, Inc. | 18,89,86,89,082 | $217.055 | $0.985 | 0.46% |
| 15 | ASML | ASML Holding N.V. | 1,96,49,48,40,000 | $494.70 | $21.73 | 4.59% |
| 16 | ATVI | Activision Blizzard, Inc | 56,97,14,77,562 | $72.825 | $0.315 | 0.43% |
| 17 | AVGO | Broadcom Inc. | 1,87,41,12,76,929 | $462.735 | $6.215 | 1.36% |
| 18 | AZN | Astrazeneca PLC | 1,73,54,50,01,728 | $56.00 | $0.82 | 1.49% |
| 19 | BIDU | Baidu, Inc. | 28,10,36,23,222 | $81.33 | $1.58 | 1.98% |
| 20 | BIIB | Biogen Inc. | 39,23,71,16,778 | $270.39 | $4.23 | -1.54% |
| 21 | BKNG | Booking Holdings Inc. | 73,62,68,98,604 | $1,854.29 | $45.73 | 2.53% |

# Dataset Description

## Nasdaq 100 Market cap.xlsx

| Variables | Description |
|---|---|
| Symbol | Symbol of the stock |
| Name | Name of the company |
| Market Cap | Market capitalization of the stock |
| Last Sale | Last sales price of the stock |
| Net Change | Current period's closing price – Previous period's closing price |
| Percentage Change | $\frac{\text{Current period's closing price – Previous period's closing price}}{\text{Previous period's closing price}} * 100$ |

# Dataset Snapshot

## nasdaq100_metrics_ratios.xlsx

| symbol | company | sector | subsector | asset_turnov | asset_turnov | asset_turnov | asset_turnov | asset_turnov |
|---|---|---|---|---|---|---|---|---|
| AAPL | Apple Inc. | Information Technology | Technology Hardware, Storage & Peripherals | 0.66 | 0.72 | 0.74 | 0.83 | 1.08 |
| ABNB | Airbnb | Consumer Discretionary | Internet & Direct Marketing Retail | | 0.55 | 0.64 | 0.36 | 0.5 |
| ADBE | Adobe Inc. | Information Technology | Application Software | 0.54 | 0.54 | 0.57 | 0.57 | 0.61 |
| ADI | Analog Devices | Information Technology | Semiconductors | 0.36 | 0.3 | 0.29 | 0.26 | 0.2 |
| ADP | ADP | Information Technology | Data Processing & Outsourced Services | | 0.34 | 0.34 | 0.35 | 0.33 |
| ADSK | Autodesk | Information Technology | Application Software | | 0.46 | 0.58 | 0.6 | 0.56 |
| AEP | American Electric Power | Utilities | Electric Utilities | 0.24 | 0.24 | 0.22 | 0.19 | 0.2 |
| ALGN | Align Technology | Health Care | Health Care Supplies | 0.93 | 1.03 | 1.06 | 0.67 | 0.73 |
| AMAT | Applied Materials | Information Technology | Semiconductor Equipment | 0.87 | 0.9 | 0.8 | 0.83 | 0.96 |
| AMD | AMD | Information Technology | Semiconductors | 1.53 | 1.6 | 1.27 | 1.3 | 1.54 |
| AMGN | Amgen | Health Care | Biotechnology | 0.29 | 0.32 | 0.37 | 0.42 | 0.42 |
| AMZN | Amazon | Consumer Discretionary | Internet & Direct Marketing Retail | 1.66 | 1.58 | 1.45 | 1.41 | 1.27 |
| ANSS | Ansys | Information Technology | Application Software | 0.38 | 0.42 | 0.37 | 0.31 | 0.31 |
| ASML | ASML Holding | Information Technology | Semiconductor Equipment | 0.54 | 0.56 | 0.55 | 0.58 | 0.63 |
| ATVI | Activision Blizzard | Communication Services | Interactive Home Entertainment | 0.39 | 0.41 | 0.34 | 0.38 | 0.37 |
| AVGO | Broadcom Inc. | Information Technology | Semiconductors | 0.34 | 0.4 | 0.38 | 0.33 | 0.36 |
| AZN | AstraZeneca | Health Care | Pharmaceuticals | 0.36 | 0.36 | 0.4 | 0.42 | 0.44 |
| BIDU | Baidu | Communication Services | Interactive Media & Services | 0.4 | 0.37 | 0.36 | 0.35 | 0.35 |
| BIIB | Biogen | Health Care | Biotechnology | 0.53 | 0.55 | 0.55 | 0.52 | 0.45 |
| BKNG | Booking Holdings | Consumer Discretionary | Internet & Direct Marketing Retail | 0.56 | 0.6 | 0.68 | 0.31 | 0.48 |

**Note**: The detailed dataset description is provided in the **Financial ratios - Data Dictionary** file under **Course Resources**

# Dataset Description

**NASDAQ_DATA**

This folder contains historical share prices for 1609 companies.
The data for 1609 companies is available in the .csv format.

**Hint:** We don't need all files given in the folder but finding relevant files can be challenging.

# Project Task: Week 1

## Applied Data Science with Python

1. Check the stock symbols of the companies in **Nasdaq 100 Market cap.xlsx**. Only the relevant files in the **NASDAQ_DATA** folder should be read.
2. Append all files (imported in the previous step) that contain no more than 10 years of data. For this, you may use your discretion.
3. Read **Nasdaq 100 market cap.xlsx** and **nasdaq100_metrics_ratios.xlsx**.
4. Collate the two files imported in the previous step to include the fields **Market cap** and **Last sale** in addition to the various metrics and ratios already present in **nasdaq100_metrics_ratios.xlsx**
5. Identify the variables whose variance is less than .005 (as these do not contribute to model building), and eliminate those variables
6. Delete the variables in **nasdaq100_metrics_ratios.xlsx** where 30% or more of the values are missing

# Project Task: Week 1

## Applied Data Science with Python

7. Perform missing value imputation for variables with less than 30% missing values by considering the company's sector
8. Analyze the effect of COVID on stock prices in detail, create visuals to support the insights, and address the following:

    a. Which sectors and companies saw the greatest impact, and which ones saw the least? You may use growth or degrowth as a measure of impact and may perform week over week, month over month (Mom), quarter over quarter (QoQ), or year over year (YoY) analysis as appropriate.

    b. Which sector and company experienced the fastest and slowest recoveries?

# Project Task: Week 1

## Machine Learning

1. Perform PCA to reduce the number of variables in the data
2. After PCA, perform cluster analysis to identify cohorts, define these cohorts (cluster profiling), and specify the insights found
3. Highlight companies from different sectors falling into the same cohort, and share your findings
4. Plot seasonality, trend, and irregular components over time for the historical stock price of Apple
5. Based on trend and seasonality, choose an appropriate exponential smoothing method to forecast the weekend share price value for the next 12 months
6. Perform an augmented Dickey–Fuller test (ADF) to check for the stationarity of Apple stock.

# Project Task: Week 1

## Machine Learning

7. Analyze the ACF and PACF plots for Apple's historical stock prices, strategize for ARIMA modeling, determine the appropriate values of p, d, and q, and forecast the month-end share price value for the next 12 months
8. Find the mean absolute percentage error (MAPE) for a 12-month period to validate the model
9. Identify the top 2 companies from each sector based on market capitalization, create trend charts for the month-end share price for the last five years (using the variable "adjusted close"), display the 12-month rolling mean and standard deviation in the same chart, and share your observations regarding the stationarity of all companies
10. Conduct an ADF test to verify the stationarity of the companies selected in the previous step
11. Perform batch forecasting for the top 2 companies from each sector based on market capitalization for the weekend share price value for the next 12 months using Auto ARIMA, and find the MAPE for a 12-month period to validate the model

**Hint:** Leverage the **pmdarima** library for Auto ARIMA

# Project Task: Week 2

## Excel

1. Create an Excel dashboard showcasing the following:

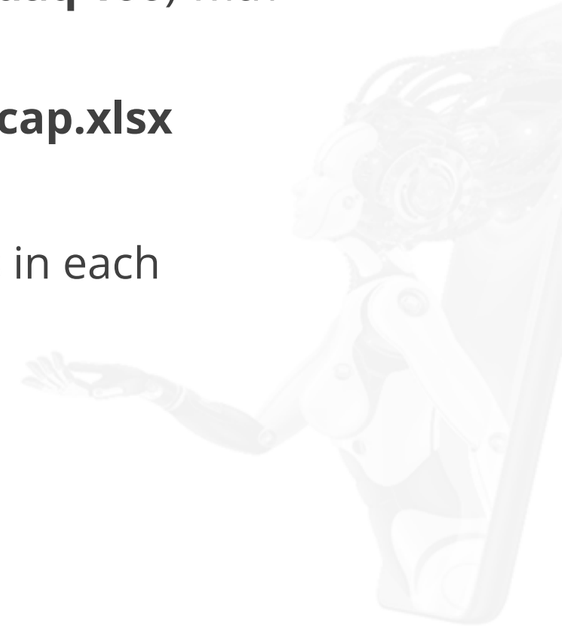
   **Hint:** Use form controls to create a dynamic chart

   a. Trend charts of share prices of the top company in each sector based on market capitalization for the past three months
   b. The market capitalization and sub industry of the companies in the dashboard
2. Create a graph to compare **yoy_revenue_growth_latest** for different companies in various industries

# Project Task: Week 2

## SQL

1. Determine the market capitalization of the company in the IT sector (from **Nasdaq 100**) with the greatest **LastSale** value
2. Find the number of companies from each industry in the **Nasdaq 100 Market cap.xlsx**
3. List the top 5 companies based on market capitalization
4. Create a table with sectors and subsectors showcasing the count of companies in each subsector



# Project Task: Week 2

## Tableau

1. Create a dashboard in Tableau by choosing appropriate chart types and metrics useful for the business

**Note:** Give emphasis to data storytelling

# Thank You