

Data Analytics

Lesson 12.

Recap and advanced topics

Dr. Hai Tran

hai.tran@sbsuni.edu.vn

Scholar: <https://scholar.google.com/citations?user=kHZvITkAAAAJ&hl=en&oi=ao>

Co-Founder: XAI - <https://xai.foo/>



**Saigon
Business
School**

In
partnership
with





Learning materials

● Textbook

- Evans, J. (2016) Business Analytics. 2nd edn. Pearson.
- Runkler, T. (2016) Data Analytics: Models and Algorithms for Intelligent Data Analysis. 2nd edn. Vieweg+Teubner Verlag.

● Online reference materials

- archive.ics.uci.edu/ml/
- powerbi.microsoft.com
- <https://github.com/topics/data-analysis-python>
- https://media.pearsoncmg.com/ph/esm/esm_evans_eba3e_20/tools/eba3e_analytic_solver.html
- <https://data.imf.org/>



Agenda

- Lesson 1: Understanding Data Analytics Terminologies.
- Lesson 2: Foundation of Business Analytics
- Lesson 3: Visualizing and Exploring data
- Lesson 4: Applying Descriptive Analytic Techniques
- Lesson 5: Data Modeling
- Lesson 6: Predictive Analytics
- Lesson 7: Regression, Classification and Clustering
- Lesson 8: Forecasting Techniques
- Lesson 9: Investigating Predictive Analytic Techniques
- Lesson 10: Introduction to Data Mining
- Lesson 11: Demonstrating Prescriptive Analytic Methods
- Lesson 12: Recap and advanced topics



Saigon
Business
School

Recap and advanced topics

s b s e d u . v n

datacamp

Python For Data Science

Data Wrangling in Pandas Cheat Sheet

Learn Data Wrangling online at [www.DataCamp.com](https://www.datacamp.com)

> Reshaping Data

Pivot

```
pd.pivot(index='date', columns='category', values='value')
```

Pivot Table

```
pd.pivot_table(values='value', index='date', columns='category')
```

Stack / Unstack

```
df.stack()
df.unstack()
```

Melt

```
pd.melt(df, id_vars='date', value_vars=['category', 'value'])
```

> Iteration

```
df.iterrows()
df.itertuples()
```

> Missing Data

```
df.isnull()
df.dropna()
df.fillna()
```

> Advanced Indexing

Also see NumPy Array

Selecting

```
df.loc[0:10, 'category']
df.iloc[0:10, 1]
```

Where

```
df[df['category'] == 'A']
```

Query

```
df.query("category == 'A'")
```

Setting/Resetting Index

```
df.set_index('date')
df.reset_index()
```

Reindexing

```
df.reindex(index=[0, 1, 2, 3])
```

Forward Filling

```
df.fillna(method='ffill')
```

Backward Filling

```
df.fillna(method='bfill')
```

Multindexing

```
df.set_index(['date', 'category'])
df.groupby(['date', 'category'])
```

> Duplicate Data

```
df.drop_duplicates()
df.duplicated()
```

> Grouping Data

Aggregation

```
df.groupby('date').agg('sum')
```

Transformation

```
df.groupby('date').transform('sum')
```

> Combining Data

Merge

```
pd.merge(df1, df2, on='date')
pd.merge(df1, df2, left_on='date', right_on='date')
```

Join

```
df1.join(df2)
```

Concatenate

```
pd.concat([df1, df2])
```

Vertical

```
df.append(df2)
```

Horizontal/Vertical

```
df1.join(df2, how='outer')
```

> Dates

```
df['date'].dt.day
df['date'].dt.month
```

> Visualization

Also see Matplotlib

```
df.plot()
```

datacamp

Descriptive Statistics

Cheat Sheet

Learn more online at [www.DataCamp.com](https://www.datacamp.com)

> Key Definitions

- Variable: In statistics, a variable is a quantity that can be measured or counted. In data analysis, a variable is a quantity in a data frame.
- Descriptive statistics: Statistics that summarize variables. They are also called summary statistics or aggregations.
- Categorical data: Data that consists of discrete groups. The categories are called labels (e.g., educational level).
- Numerical data: Data that consists of numbers (e.g., age).

> Numerical Dataset—Glasses of Water

Visualizing Numeric Variables

There are a variety of ways of visualizing numerical data, each a form of data in action.

Histogram

Shows the distribution of a variable. It plots numerical data into bins or columns. The x-axis shows the range, and the y-axis represents the frequency.

Box plot

Shows the distribution of a variable using a box summarizing statistics—minimum, first quartile, median, third quartile, and maximum.

Measures of Center

Measure	Definition	How to Find It	Result
Arithmetic mean	The total of the values divided by how many values there are	$\frac{100 + 100 + 100 + 100 + 100}{5}$	100.0
Median	The middle value when values are ordered from smallest to largest	100	100.0
Mode	The most common value	100	100.0

Other Measures of Location

Measure	Definition	How to Find It	Result
Minimum	The lowest value in your data	100	100.0
Maximum	The highest value in your data	100	100.0

Measures of Spread

Measure	Definition	How to Find It	Result
Range	The highest value minus the lowest value	$100 - 100$	0.0
Variance	The sum of the squares of the difference between each value and the mean, divided by one less than the number of data points	$\frac{(100 - 100)^2 + (100 - 100)^2 + (100 - 100)^2 + (100 - 100)^2 + (100 - 100)^2}{5 - 1}$	0.0
Inter-quartile range	The third quartile minus the first quartile	100 - 100	0.0

> Correlation

Correlation is a measure of the linear relationship between two variables. That is, when one variable goes up, does the other variable go up or down? There are several algorithms to calculate correlation, but it is always a score between -1 and 1.

Correlation score	Interpretation
1	When X increases, Y increases. Scatter plot forms a perfect straight line with positive slope.
-1	When X increases, Y decreases. Scatter plot forms a perfect straight line with negative slope.
0	There is no linear relationship between X and Y, as the scatter plot looks like a noisy mess.
Between 0 and 1	When X increases, Y increases. Scatter plot forms a perfect straight line with positive slope.
Between -1 and 0	When X increases, Y decreases. Scatter plot forms a perfect straight line with negative slope.

Note that correlation does not account for non-linear effects, so if X and Y do not have a straight line relationship, the correlation score may not be meaningful.

> Categorical Data—Trail Mix

To illustrate statistical concepts on categorical data, we'll use an unsorted categorical variable, consisting of several elements of a trail mix. Our categorical variable contains 10 elements: 10 almonds, 10 cashews, and 10 raisins.

Counts and Proportions

Food category	Count	Proportion
Almond	10	$\frac{10}{40} = 0.25$
Cashew	10	$\frac{10}{40} = 0.25$
Raisin	10	$\frac{10}{40} = 0.25$

Visualizing Categorical Variables

Bar plot, Stacked bar chart, Treemap chart



Recap and advanced topics

The Data Visualization Cheat Sheet

Learn Data Visualization online at [www.DataCamp.com](https://www.datacamp.com)

How to use this cheat sheet

Use this cheat sheet for inspiration when making your next data visualizations. For more data visualization cheat sheets, check out our cheat sheets repository [here](#).

Part-to-whole charts

Pie chart

Use a pie chart to show the relative size of different categories in a whole. It's best used for a small number of categories.

USE CASES

- 1. Show the relative size of different categories in a whole.
- 2. Show the relative size of different categories in a whole.

Donut chart

The donut chart is a variation of the pie chart, but it's better for showing a single data point in the center for readability.

USE CASES

- 1. Highlight a single data point in the center of the chart.
- 2. Highlight a single data point in the center of the chart.

Heat maps

Heat maps are a way to visualize data that's too complex to show in a table. They use color to represent the magnitude of the data.

USE CASES

- 1. Visualize data that's too complex to show in a table.
- 2. Visualize data that's too complex to show in a table.

Stacked column chart

Use a stacked column chart to show the relative size of different categories in a whole. It's best used for a small number of categories.

USE CASES

- 1. Show the relative size of different categories in a whole.
- 2. Show the relative size of different categories in a whole.

Treemap chart

Use a treemap chart to show the relative size of different categories in a whole. It's best used for a small number of categories.

USE CASES

- 1. Show the relative size of different categories in a whole.
- 2. Show the relative size of different categories in a whole.

Capture a trend

Line chart

The most straightforward way to show a trend over time is with a line chart. It's best used for a small number of categories.

USE CASES

- 1. Show a trend over time.
- 2. Show a trend over time.

Area chart

Use an area chart to show a trend over time. It's best used for a small number of categories.

USE CASES

- 1. Show a trend over time.
- 2. Show a trend over time.

Bar chart

Use a bar chart to show a trend over time. It's best used for a small number of categories.

USE CASES

- 1. Show a trend over time.
- 2. Show a trend over time.

Stacked area chart

Use a stacked area chart to show a trend over time. It's best used for a small number of categories.

USE CASES

- 1. Show a trend over time.
- 2. Show a trend over time.

Bubble chart

Use a bubble chart to show a trend over time. It's best used for a small number of categories.

USE CASES

- 1. Show a trend over time.
- 2. Show a trend over time.

Visualize a single value

Gauge

Use a gauge to show a single value. It's best used for a small number of categories.

USE CASES

- 1. Show a single value.
- 2. Show a single value.

Table chart

Use a table chart to show a single value. It's best used for a small number of categories.

USE CASES

- 1. Show a single value.
- 2. Show a single value.

Gauge chart

Use a gauge chart to show a single value. It's best used for a small number of categories.

USE CASES

- 1. Show a single value.
- 2. Show a single value.

Capture distributions

Histogram

Use a histogram to show a distribution of data. It's best used for a small number of categories.

USE CASES

- 1. Show a distribution of data.
- 2. Show a distribution of data.

Box plot

Use a box plot to show a distribution of data. It's best used for a small number of categories.

USE CASES

- 1. Show a distribution of data.
- 2. Show a distribution of data.

Violin plot

Use a violin plot to show a distribution of data. It's best used for a small number of categories.

USE CASES

- 1. Show a distribution of data.
- 2. Show a distribution of data.

Density plot

Use a density plot to show a distribution of data. It's best used for a small number of categories.

USE CASES

- 1. Show a distribution of data.
- 2. Show a distribution of data.

Visualize relationships

Bar chart

Use a bar chart to show a relationship between two variables. It's best used for a small number of categories.

USE CASES

- 1. Show a relationship between two variables.
- 2. Show a relationship between two variables.

Scatter plot

Use a scatter plot to show a relationship between two variables. It's best used for a small number of categories.

USE CASES

- 1. Show a relationship between two variables.
- 2. Show a relationship between two variables.

Correlation coefficient

Use a correlation coefficient to show a relationship between two variables. It's best used for a small number of categories.

USE CASES

- 1. Show a relationship between two variables.
- 2. Show a relationship between two variables.

Bubble chart

Use a bubble chart to show a relationship between two variables. It's best used for a small number of categories.

USE CASES

- 1. Show a relationship between two variables.
- 2. Show a relationship between two variables.

Word cloud chart

Use a word cloud chart to show a relationship between two variables. It's best used for a small number of categories.

USE CASES

- 1. Show a relationship between two variables.
- 2. Show a relationship between two variables.

Visualize a flow

Sankey diagram

Use a Sankey diagram to show a flow of data. It's best used for a small number of categories.

USE CASES

- 1. Show a flow of data.
- 2. Show a flow of data.

Chord diagram

Use a chord diagram to show a flow of data. It's best used for a small number of categories.

USE CASES

- 1. Show a flow of data.
- 2. Show a flow of data.

Network diagram

Use a network diagram to show a flow of data. It's best used for a small number of categories.

USE CASES

- 1. Show a flow of data.
- 2. Show a flow of data.

Learn Data Skills Online at [www.DataCamp.com](https://www.datacamp.com)

Data Storytelling & Communication Cheat Sheet

Learn more about data storytelling at [www.DataCamp.com](https://www.datacamp.com)

What is data storytelling?

Data storytelling is a way to communicate data in a way that's easy to understand. It's best used for a small number of categories.

Crafting effective visuals

Choose the best visualization for your story. It's best used for a small number of categories.

Crafting effective narratives with data

Know the audience. It's best used for a small number of categories.

Use text appropriately

Use text to highlight key points in your data. It's best used for a small number of categories.

Use color effectively

Use color to highlight key points in your data. It's best used for a small number of categories.

Do not mislead with data stories

Do not mislead with data stories. It's best used for a small number of categories.

Crafting effective narratives with data

Know the audience. It's best used for a small number of categories.

Choose the best medium to share your story

Choose the best medium to share your story. It's best used for a small number of categories.

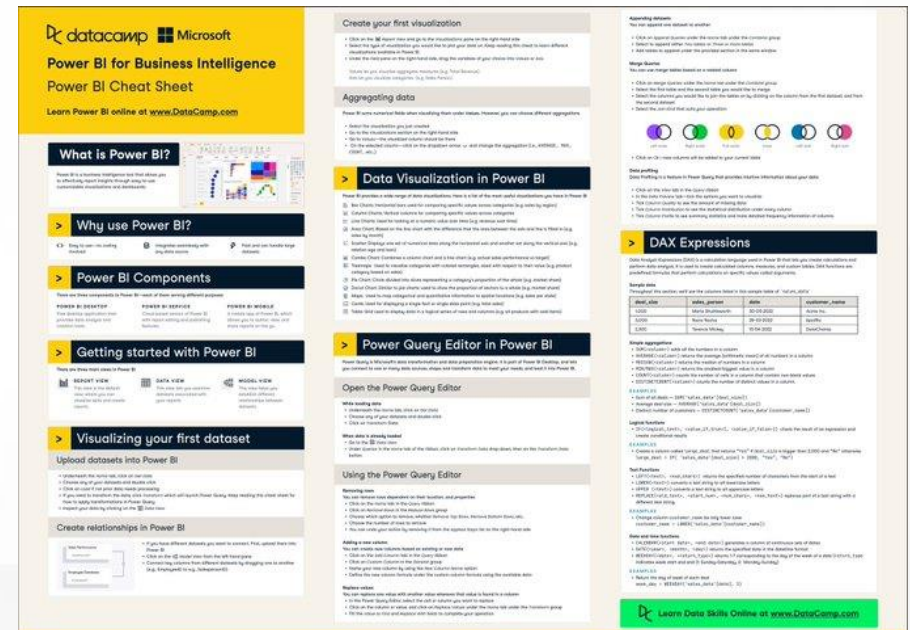
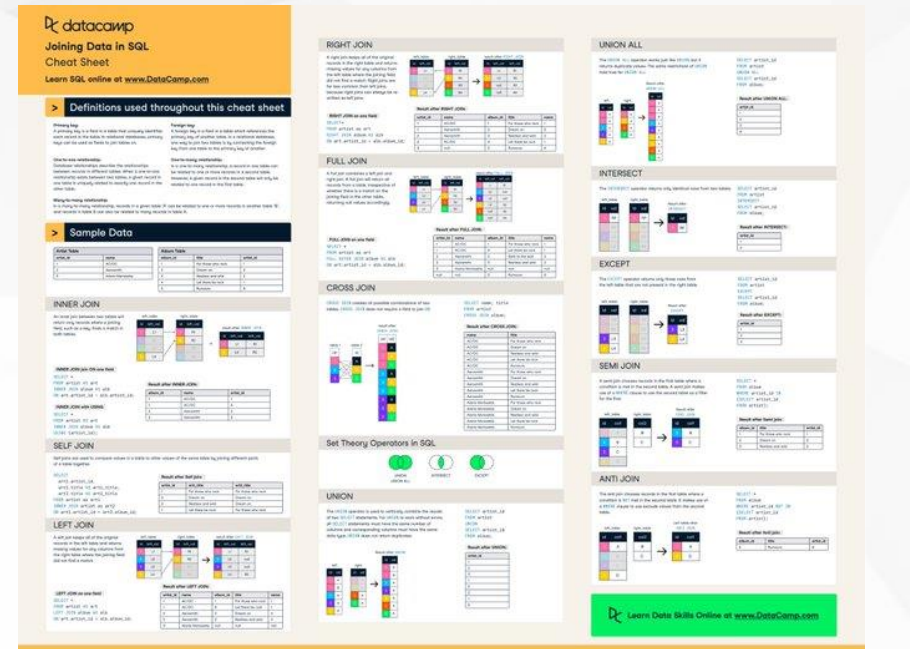


Saigon
Business
School

Conclusion & Questions

Here are five advanced topics in data analytics:

- 1. Deep Learning and Neural Networks:** Explore advanced techniques in deep learning, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to enhance pattern recognition and prediction capabilities.
- 2. Natural Language Processing (NLP) and Text Analytics:** Delve into the complexities of analyzing and understanding human language, including sentiment analysis, text summarization, and language generation, to derive insights from unstructured text data.
- 3. Time Series Analysis and Forecasting:** Focus on analyzing data that varies over time, such as stock prices, weather patterns, or sales figures, using sophisticated methods like ARIMA (AutoRegressive Integrated Moving Average) and machine learning models for accurate forecasting.
- 4. Geospatial Analytics:** Explore the integration of geographic information systems (GIS) with data analytics, allowing for the analysis and visualization of spatial patterns, location-based insights, and geospatial data to make informed decisions.
- 5. Explainable AI (XAI):** Address the interpretability and transparency challenges associated with complex machine learning models, ensuring that the decisions made by these models can be understood and trusted by stakeholders, especially in sensitive domains like healthcare and finance.





Saigon
Business
School

Thank you

s b s e d u . v n

