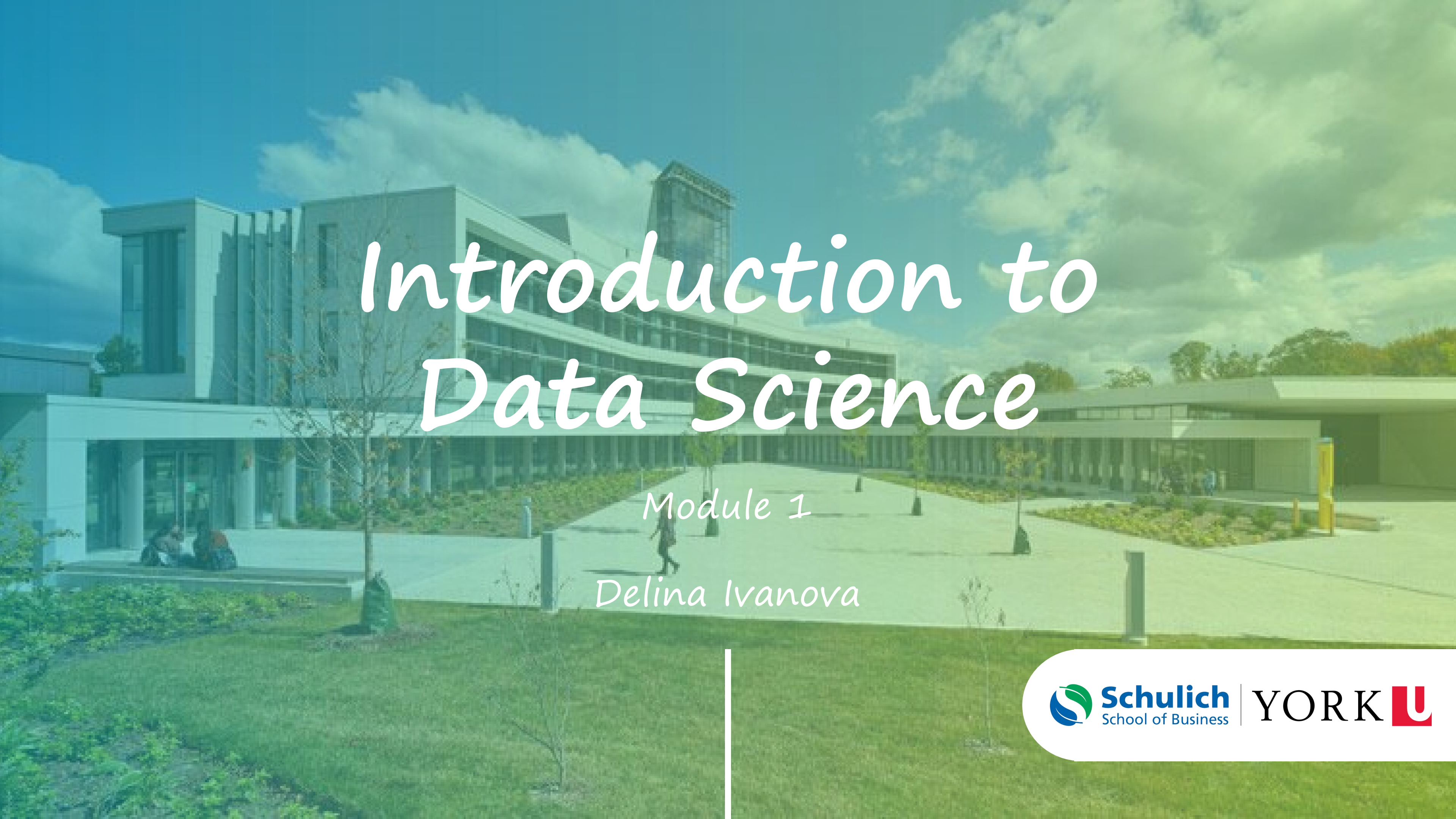


Introduction to Data Science

Module 1

Delina Ivanova

Schulich School of Business | YORK U

About Me

- 10+ years of experience in analytics across mobile gaming, CPG, banking, and public sector

- Currently:
  - Director of Analytics @ **Mistplay** (mobile gaming loyalty platform) leading analytics teams partnering with Product, Marketing, LiveOps, Loyalty and Commercial teams
  - Data Science and Machine Learning instructor @Schulich, University of Toronto and University of Waterloo

- Previously:
  - Associate Director, Data & Analytics @ HelloFresh Canada (meal kit delivery)
  - Senior Manager, Analytics @ TD Bank
  - Manager, Analytics and Data Innovation @ TD Bank

Course Objectives

Throughout the course, you will learn to:

- Conduct exploratory data analysis and data cleaning using Python libraries such as Pandas, NumPy, SciPy, Matplotlib, and Seaborn
- Apply descriptive statistics, probability theory, and statistical inference techniques to analyze and interpret data
- Design and evaluate A/B tests and experiments for real-world applications
- Perform feature engineering, including dimensionality reduction, variable transformations, and feature scaling
- Build, evaluate, and optimize various machine learning models, such as linear and logistic regression, naive bayes, k-nearest neighbors, support vector machines, decision trees, and random forests

Module 1

Exploratory Data Analysis
and Data Cleaning

1. Introduction to Data Science
2. Introduction to Exploratory Data Analysis (EDA)
3. Introduction to Data Cleaning
4. Python Libraries for EDA and Data Cleaning
5. EDA Techniques
6. Data Cleaning Techniques

Introduction to Data Science

What is data science?

✓ Data Science Is...

01 A toolbox to derive an understanding from data

02 An enabler for better decision making

03 A probabilistic method for solving complex problems

Data Science Is Not... ✗

01 Magic

02 A crystal ball

03 100% correct all the time

Data Science and Analytics in Organizations

- There are various types of data teams within an organization:
    - Data platform: teams who build and maintain the infrastructure to enable analytics, experimentation, modelling and AI
    - Data engineering: teams who build pipelines to ingest data from external sources in a central location
    - Data analytics: teams who provide reporting, analysis, and experimentation support to business partners
    - Data science: teams who build decision models (e.g., media mix, propensity), or user-facing AI products
    - Data governance: teams who build processes to effectively manage data collection and use

Types of Business Problems

- Revenue Generation
  - Product analytics and customer preferences
  - Marketing analytics

- Cost Reduction
  - Product mix optimization
  - Production / logistics / supply chain optimization

- Customer Experience
  - Product recommendation
  - Intelligent chat bots

- Risk Management
  - Fraud detection
  - Expense management / loss prevention

Types of Data Science Solutions

- Descriptive
  - Answers the question of "What happened?"
  - Reporting-based, descriptive statistics
- Diagnostic
  - Answers the question of "Why did it happen?"
  - Uses inferential statistics to draw statistically significant conclusions
- Predictive
  - Answers the question of "What could happen?"
  - Uses modelling techniques to predict potential outcomes based on a series of inputs
- Prescriptive
  - Answers the question of "What should happen?"
  - Uses modelling techniques to identify the most optimal solution within a set of constraints

Exploratory Data Analysis

Introduction to Exploratory Data Analysis

- **Exploratory Data Analysis (EDA) is the process of analyzing and visualizing data to extract insights and identify patterns, trends, and relationships between data inputs and an output.**

- **EDA Helps to understand the data, generate hypotheses, identify potential issues, and inform subsequent analysis steps.**

- **It helps us uncover the underlying structure of the data, detect outliers and anomalies, test assumptions, and develop**

The Data Exploration Process

1. Define the business problem – what are we trying to understand?
2. Identify the data required to solve the problem
3. Create a data set
    - Internal data (SQL)
    - External data (csv files, APIs)
4. Conduct univariate analysis (each variable on its own)
5. Conduct bi-variate and multi-variate analysis
6. Identify meaningful relationships
7. Interpret the results and develop

Data Cleaning

Introduction to Data Cleaning

- Data should always be cleaned to ensure accurate and reliable analysis, improve data quality, and reduce the risk of errors.
- Common Data Quality Issues:
  - Missing values
  - Duplicates
  - Data entry errors
  - Incorrect data types
  - Outliers

The Data Cleaning Process

1. Use visual inspection, summary statistics and specific investigations to identify data issues.
2. Develop a cleaning strategy, considering the pros and cons of different cleaning techniques.
3. Clean the data.
4. Validate the results; adjust as required.

Python Libraries

Python Libraries for Data Cleaning and EDA

- Pandas: A powerful library for data manipulation, cleaning, and analysis.
- NumPy: A library for numerical computing, including support for arrays and mathematical functions.
- SciPy: A library for scientific computing, including optimization, linear algebra, integration, interpolation, signal and image processing, and advanced statistical functions.
- Matplotlib: A library for creating static, interactive, and animated

Data Exploration Techniques

Univariate Analysis

Descriptive Statistics

- Descriptive Statistics summarize and describe the main features of a dataset using numerical measures and visualizations.
- Measures of Central Tendency are metrics which provide a single value that represents the center of the data distribution.
  - Mean: The average of all data points, calculated as the sum of the values divided by the number of values (sensitive to extreme values/outliers).
  - Median: The middle value in a dataset when the data points are sorted in

Measures of Central Tendency

Measure	**Description**	**Formula**
Mean	**The average of all data points, calculated as the sum of the values divided by the number of values (sensitive to extreme values/outliers).**	$\overline{X} = \frac{\sum X}{N}$ $\mathrm{Med}(X)$ $\begin{cases} X[\frac{n+1}{2}] & \text{if n is odd} \\ \frac{X[\frac{n}{2}]+X[\frac{n}{2}+1]}{2} & \text{if n is even} \end{cases}$
Median	**The middle value in a dataset when the data points are sorted in ascending or descending order (less sensitive to extreme values/outliers).**	

Measures of Dispersion

1. **Measures of Dispersion: Describe the spread or variability of the data distribution.**
   1. **Range: The difference between the maximum and minimum values in the dataset (sensitive to extreme values/outliers).**
   2. **Variance: The average of the squared differences from the mean, indicating the dispersion of the data points (larger values indicate more spread).**
   3. **Standard Deviation: The square root of the variance, representing the average deviation from the mean (easier to interpret as it is in the same unit as the data).**
   4. **Interquartile Range (IQR): The difference between the first quartile (25th percentile)**

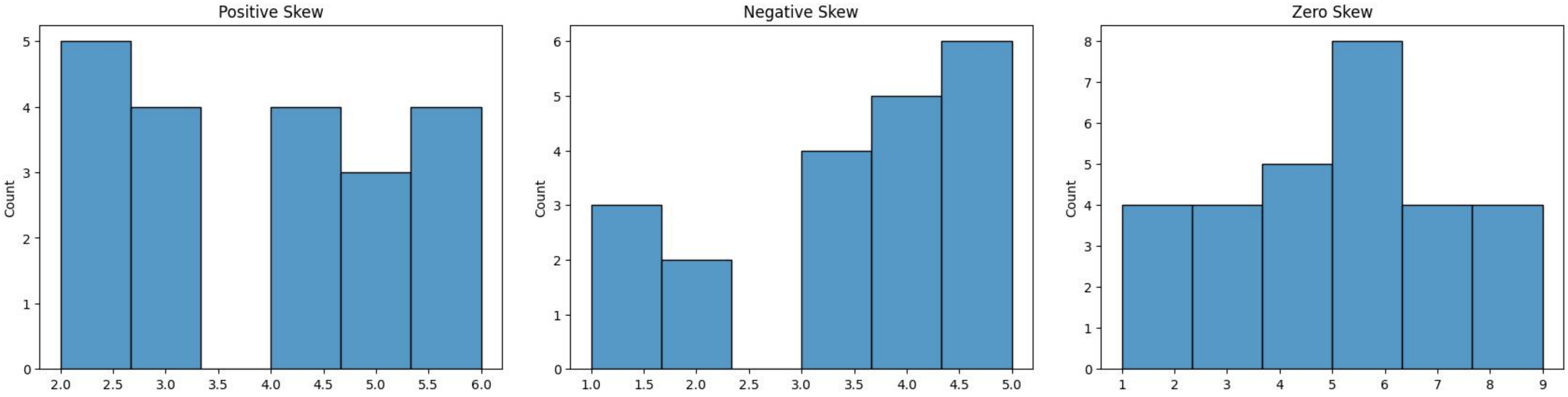
Measures of Dispersion

Measure	**Description**	**Formula**
Range	**The difference between the maximum and minimum values in the dataset (sensitive to extreme values/outliers).**	$\text{Range}(X) = \text{Max}(X) - \text{Min}(X)$ $S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
Variance	**The average of the squared differences from the mean, indicating the dispersion of the data points (larger values indicate more spread).**	$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$ $IQR = Q_3 - Q_1$
Standard Deviation	**The square root of the variance, representing the average deviation from the mean (easier to interpret as it is in the same unit as**	

Skewness and Kurtosis

1. **Skewness: A measure of the asymmetry of the data distribution, indicating whether the data is skewed to the left (negatively skewed) or right (positively skewed).**
2. **Kurtosis: A measure of the "tailedness" of the data distribution, indicating whether the data has heavy tails (high kurtosis) or light**

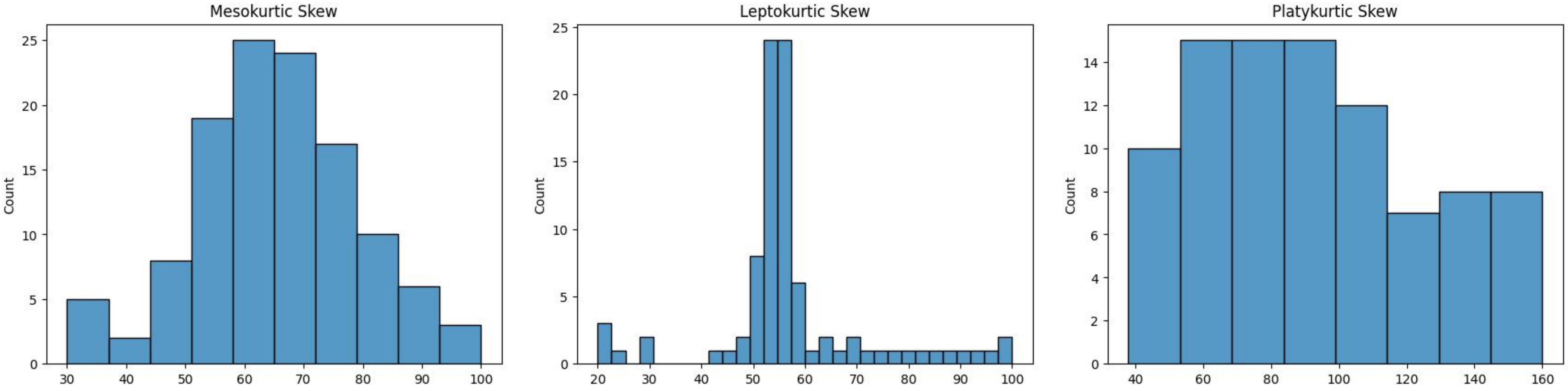
Skewness


Positive Skew
Count
0
1
2
3
4
5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Negative Skew
Count
0
1
2
3
4
5
6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Zero Skew
Count
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
9

Examples of Skewness

- **Income distribution: If most people have low incomes and only a few have high incomes, the income distribution would be positively skewed. The mean would be higher than the median and mode.**
- **Retirement: If most people retire around the age of 65, but some people retire very early around the age of 40-45, the**

Kurtosis


Mesokurtic Skew
Count
0
5
10
15
20
25
30
40
50
60
70
80
90
100
Leptokurtic Skew
Count
0
5
10
15
20
25
20
30
40
50
60
70
80
90
100
Platykurtic Skew
Count
0
2
4
6
8
10
12
14
40
60
80
100
120
140
160

Examples of Kurtosis

- **Height of adult humans: If most people in a population are around 5’8”, and few people are taller or shorter, the distribution would be mesokurtic (normally distributed).**
- **Test scores: If a very easy test was given to students, most students would score well, making the test score distribution platykurtic (uniform, wide and flat).**
- **High-risk investment returns: If an investment has extreme increases and decreases, the distribution would**

Calculating Skewness and Kurtosis

Measure	Description	Formula
Skewness	A measure of the asymmetry of the data distribution, indicating whether the data is skewed to the left (negatively skewed) or right (positively skewed). Skew = 0 indicates no skew; >0 indicates right (or positive skew); <0 indicates left (or negative skew).	$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$ $\text{Kurtosis} = \frac{\sum \left(x - \bar{x}\right)^4}{(n - 1) \cdot S^4}$
	A measure of the "tailedness" of the data distribution, indicating whether the data has heavy tails (high kurtosis) or	

Python Demo

1. **Open a Jupyter Notebook in VS Code, and follow along with the demo.**

In-class Exercise – Descriptive Statistics

1. Suppose you had the following sample data representing incomes from individuals, in thousands.

   data = [45, 55, 60, 50, 80, 62, 54, 58, 71, 48]

   Calculate descriptive statistics using NumPy and Scipy, including:

   a. Measures of central tendency
   b. Measures of dispersion
   c. Skewness
   d. Kurtosis

Data Visualization

Visualizations

1. Histogram
2. Boxplot

Python Demo

1. **Open a Jupyter Notebook in VS Code, and follow along with the demo.**

Data Cleaning Techniques

Missing Values

- Strategies for Handling Missing Values:
  - Listwise deletion
  - Pairwise deletion
  - Imputation
  - Interpolation
- Imputation Techniques:
  - Mean
  - Median
  - Mode imputation
  - Regression imputation
  - K-nearest neighbours imputation.

Removing Duplicates

- Identifying Duplicate Records: Use Pandas to find rows with identical values in specified columns.
- Removing Duplicates Using Pandas: Use the drop_duplicates() function to remove duplicate rows.

Data Entry Errors

- Common Types of Data Entry Errors: Typos, misspellings, inconsistent formatting.
- Data Validation Techniques: Regular expressions, lookup tables, data dictionaries.
- Manual Correction vs. Automated Correction: Manual correction is time-consuming and error-prone; automated correction using scripts and tools can be faster and more reliable.

Converting Data Types

- Common Data Type Conversions: String to numeric, date to string, string to date.
- Converting Data Types Using Pandas: Use the astype() function to change data types.

Data Distributions

- Normal Distribution (Gaussian Distribution): A continuous probability distribution that is symmetric about the mean and characterized by its bell shape.
  - Mean (μ): The center of the distribution
  - Standard Deviation (σ): The spread of the distribution
- Standard Normal Distribution: A special case of the normal distribution where $\mu = 0$ and $\sigma = 1$. The Z-score represents how many standard deviations a data point is from the mean of the standard normal distribution.
  - Z-score: $Z = (X - \mu) / \sigma$
    - Z: Z-score
    - X: data point
    - μ: mean of the distribution
    - σ: standard deviation of the distribution
- Importance of Data Distributions:
  - Understanding the distribution of the data helps us make better assumptions and select appropriate statistical techniques.
  - Many statistical tests and machine learning algorithms assume that the data follows a normal distribution.

Outliers

- Identifying Outliers: Use visualizations (e.g., box plots, histograms) or statistical methods (e.g., Z-score, IQR) to detect potential outliers.
- Strategies for Handling Outliers: Remove, transform, or cap/floor outliers, depending on the context and the potential impact on the analysis.

1.

In-class Exercise – Data Cleaning

1. Using the provided "messy_data.csv" file:
   - In groups or individually, have students identify and correct the data quality issues using Python libraries (e.g., Pandas).
   - Encourage students to share their solutions and discuss any challenges they encountered during the exercise.

Additional Resources

- Recap of Week 1 Topics: EDA, data cleaning, Python libraries, EDA techniques, and data cleaning tasks.
- Importance of EDA and Data Cleaning in Data Science: Ensures reliable and accurate analysis, uncovers insights, and sets the stage for subsequent steps in the data science process.

1.

Thank you!

See you next
week.