

Introduction to Data Science and Machine Learning

**VERSIONING WITH GIT AND DATA
PREPARATION (PART 2)**

- **Addition of Teams**
- **Discussion of Tasks**
- **Introduction to Git – Part 2**
- **Additional Comments on the Feature Engineering**
- **Introduction to Analyzing Time Series Data**

The background of the slide is a dark, almost black, surface with intricate, swirling patterns resembling liquid or smoke. These patterns are composed of various shades of dark grey and black, creating a sense of depth and movement.

ADDITION OF TEAMS

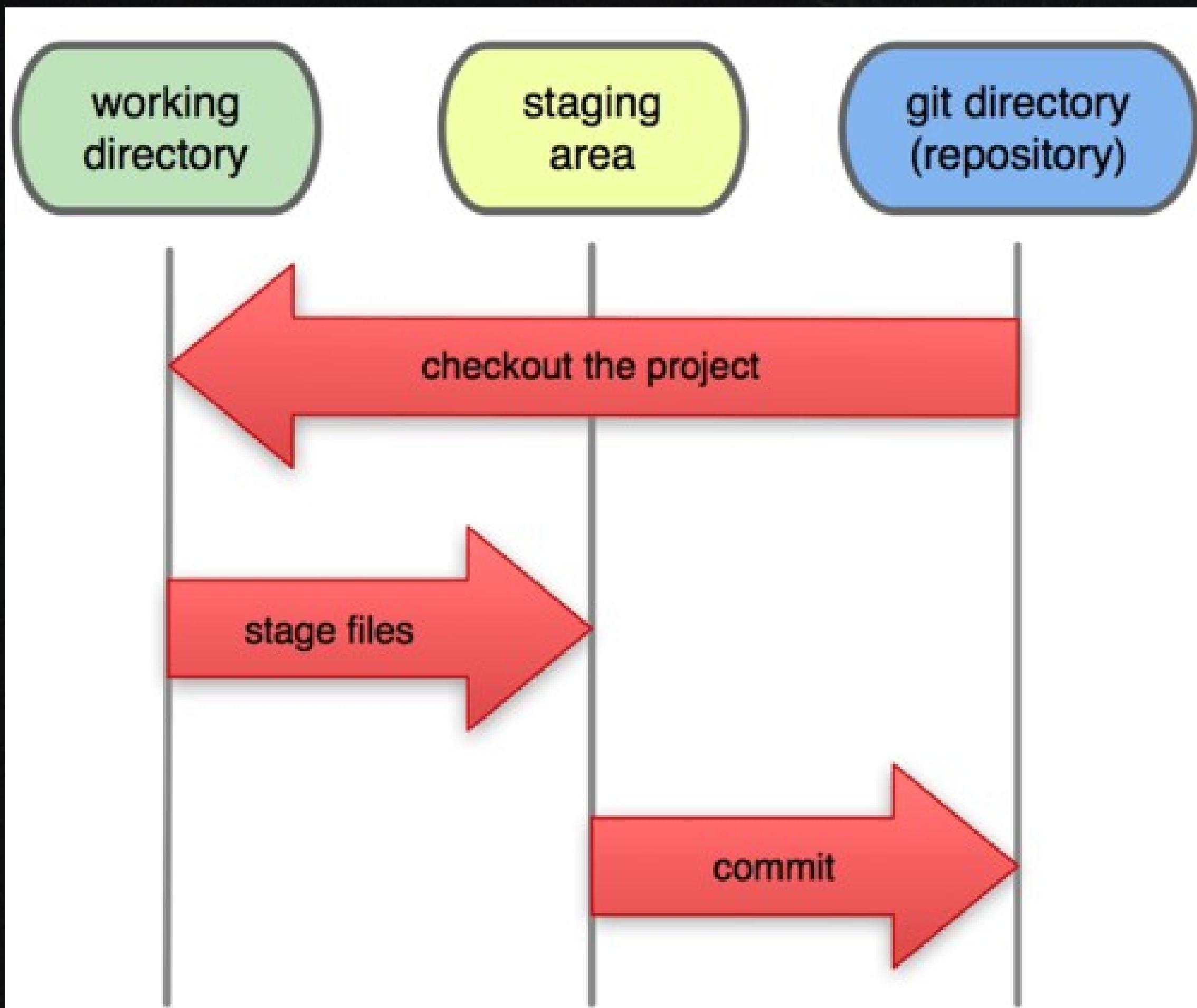
BREAKOUT

- If necessary, a brief round of introductions again
- Compare which methods you used to merge the datasets
- Present the results of the descriptive statistics and visualizations you created

WEATHER CODE

- Which visualization or analysis could be helpful here?

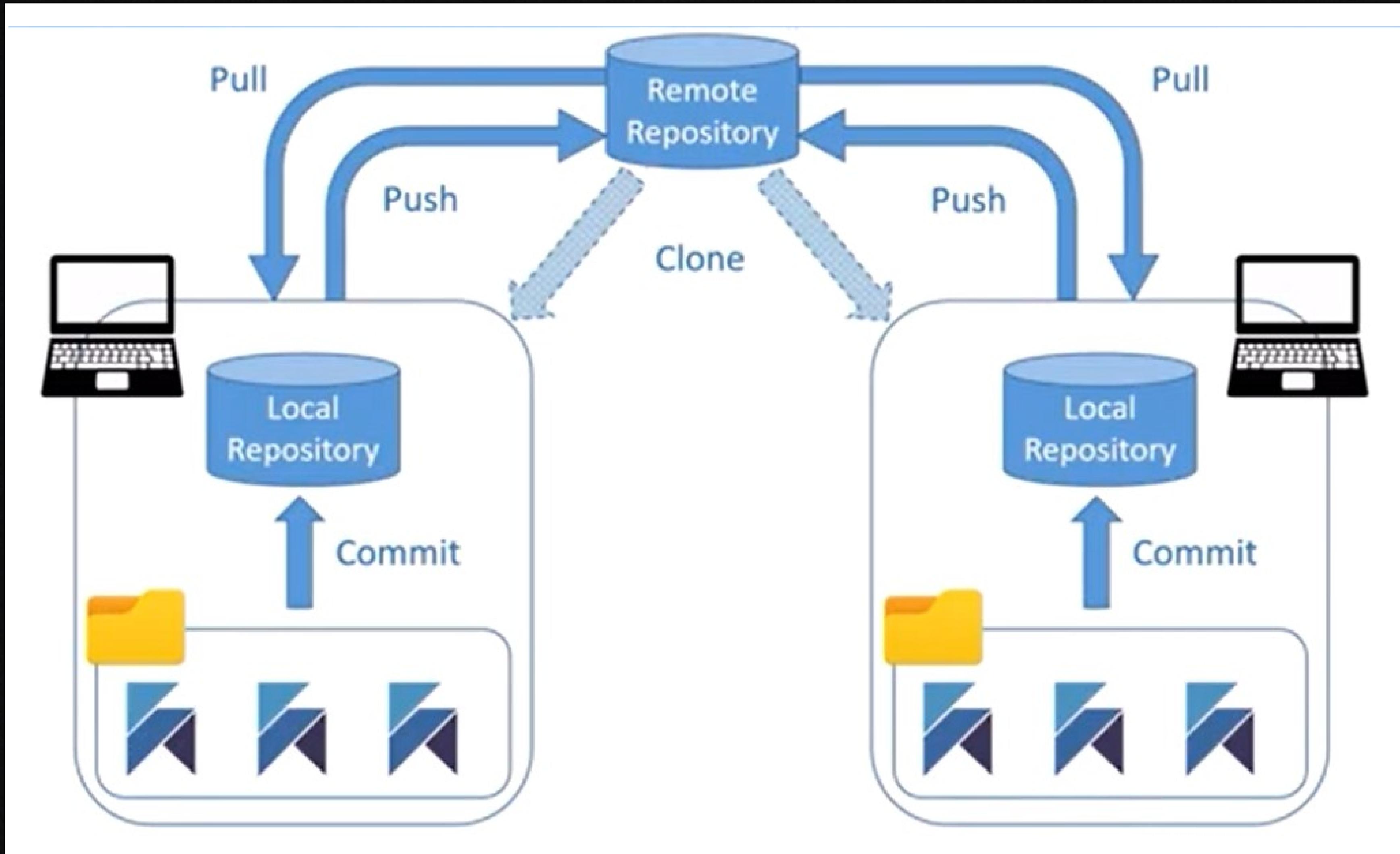
VERSIONING WITH GIT



A file can have the following states:

- Ignored (excluded from version control via `.gitignore`)
- Untracked (not yet under version control)
- Modified (changed compared to the last versioned file)
- Staged (marked for inclusion in the next version)
- Committed (under version control)

LOCAL AND REMOTE REPOSITORIES



AUTHENTICATION

- You or the application on your computer need read and write permissions on the remote server.
- To enable reading and writing, a personal key is typically used—one that is known only to you (or your computer) and the remote server.
- This key allows the remote server to identify you and verify your permissions on the server.

AUTHENTICATION WITH GITHUB

If you use GitHub Codespaces:

- Authentication is handled automatically via your GitHub account.

If you use a local VS Code:

- After installing the VS Code extension “*Github Pull Requests and Issues*”, you’ll be automatically prompted to log in to GitHub. During this process, a private key is generated for you by GitHub and stored locally on your machine.

CREATING THE TEAM REPOSITORY

The screenshot shows a GitHub repository page for 'repo-template-intro-to-data-science-and-ml'. The repository is owned by 'opencampus-sh'. The main navigation bar includes links for Code, Issues, Pull requests, Actions, Security, Insights, and Settings. The 'Code' tab is selected. The repository has 1 branch and 0 tags. A search bar at the top right allows users to search for files. The repository's description is 'Repository Template für den Kurs Einführung in Data Science und maschinelles Lernen'. It has 2 stars, 4 watchers, 1 fork, and 5 commits. The commit history lists the following changes:

File / Commit	Description	Date
steffen74 Update README.md	updates instructions	c6965ea · 7 months ago
0_DataPreparation	updates instructions	last year
1_DatasetCharacteristics	updates instructions	last year
2_BaselineModel	updates instructions	last year
3_Model	updates instructions	last year
4_Presentation	initial version	last year
CoverImage	initial version	last year
.gitignore	Initial commit	last year
README.md	Update README.md	7 months ago
repo-template-intro-to-data-science-an...	initial version	last year

The 'About' section provides details such as Readme, Activity, Custom properties, 2 stars, 4 watching, 1 fork, and a link to Report repository. The 'Releases' section indicates no releases have been published, with a link to Create a new release. The 'Packages' section also indicates no packages have been published, with a link to Publish your first package.

https://github.com/new?template_name=repo-template-intro-to-data-science-and-ml&template_owner=opencampus-sh



New repository

Type / to search



Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere?
[Import a repository.](#)

Required fields are marked with an asterisk (*).

Repository template

[opencampus-sh/repo-template-intro-to-data-science-and-ml](#) ▾

Start your repository with a template repository's contents.

Include all branches

Copy all branches from opencampus-sh/repo-template-intro-to-data-science-and-ml and not just the default branch.

Owner *



steffen74 ▾

Repository name *



/



Great repository names are short and memorable. Need inspiration? How about [effective-fishstick](#) ?

Description (optional)

Public

Anyone on the internet can see this repository. You choose who can commit.

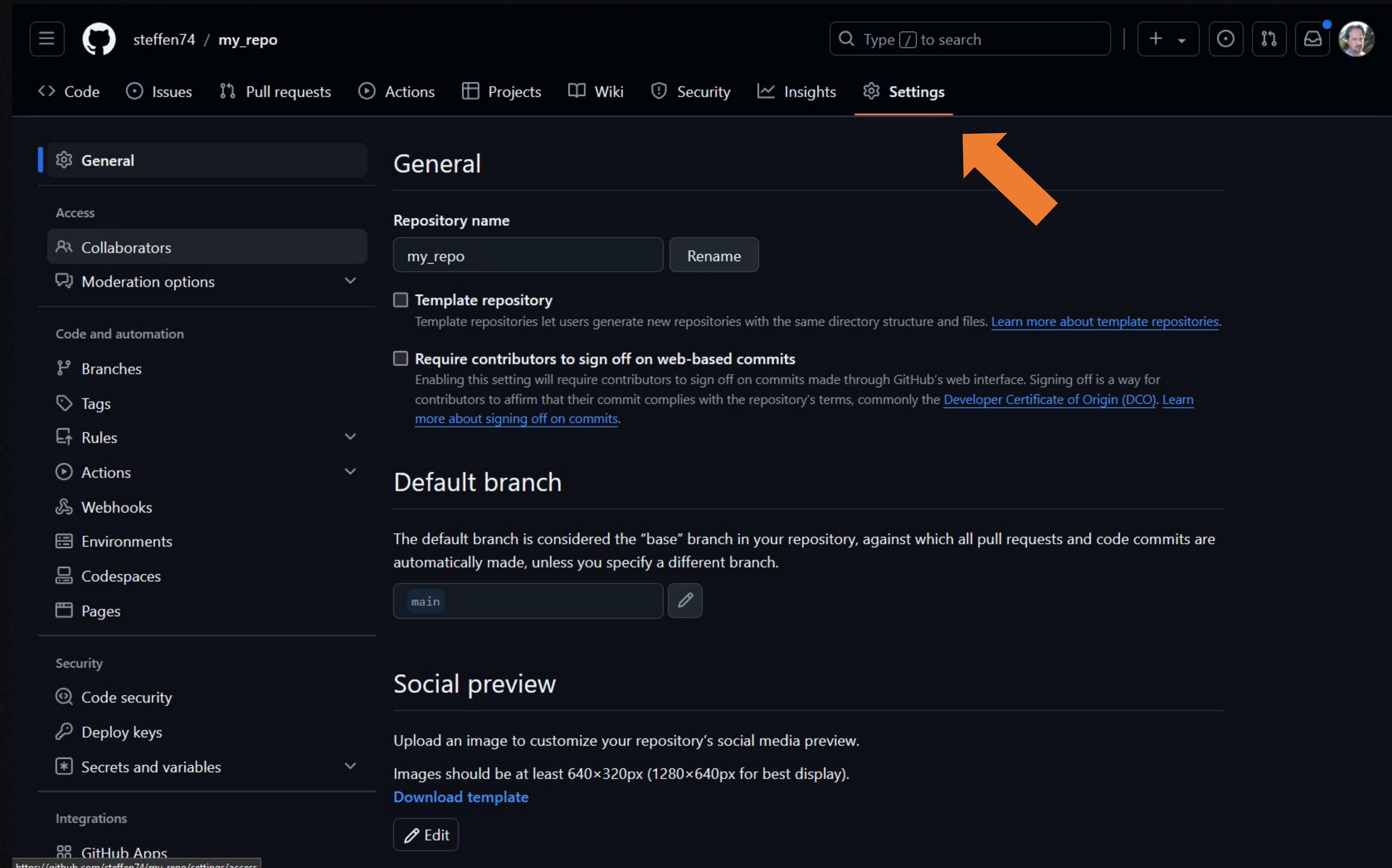
Private

You choose who can see and commit to this repository.

You are creating a public repository in your personal account.



Create repository





steffen74 / my_repo



Type / to search

[Code](#) [Issues](#) [Pull requests](#)[Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#) [Settings](#)

General

Access

Collaborators

Moderation options

Code and automation

Branches

Tags

Rules

Actions

Webhooks

Environments

Codespaces

Pages

Security

Code security

Deploy keys

Secrets and variables

Integrations

GitHub Apps

Add a collaborator to my_repo

Search by username, full name, or email

 Find people[Manage](#)

Select a collaborator above

[Manage](#)

to this repository.

Manage access



You haven't invited any collaborators yet

[Add people](#)

BREAKOUT

- *One person in the team creates the team repository.*
- All team members create a GitHub Codespace based on this repo.
- One person copies their code for merging the data into their Codespace for the team repo and pushes the code, making it available for everyone else.
- The remaining team members pull the newly added code.
- Each of you should also push the code for your descriptive statistics and visualizations they made.

DATA PREPARATION FOR MACHINE LEARNING

- **Data Collection**

Collecting raw data from various sources such as databases, files, APIs, or web scraping.

- **Cleaning**

Handling missing data, removing duplicates, and correcting errors.

- **Data Exploration / Analysis**

Understanding the distribution of variables, identifying outliers and relationships—for example, using statistical metrics, correlation matrices, or visualization tools.

- **Feature-Engineering**

Transforming raw data into features that better represent the underlying problem for models (e.g., creating new variables from existing data), encoding categorical variables, and handling missing values.

- **Data Splitting**

Dividing the data into training, validation, and test sets to ensure that the model generalizes well to new, unseen data..

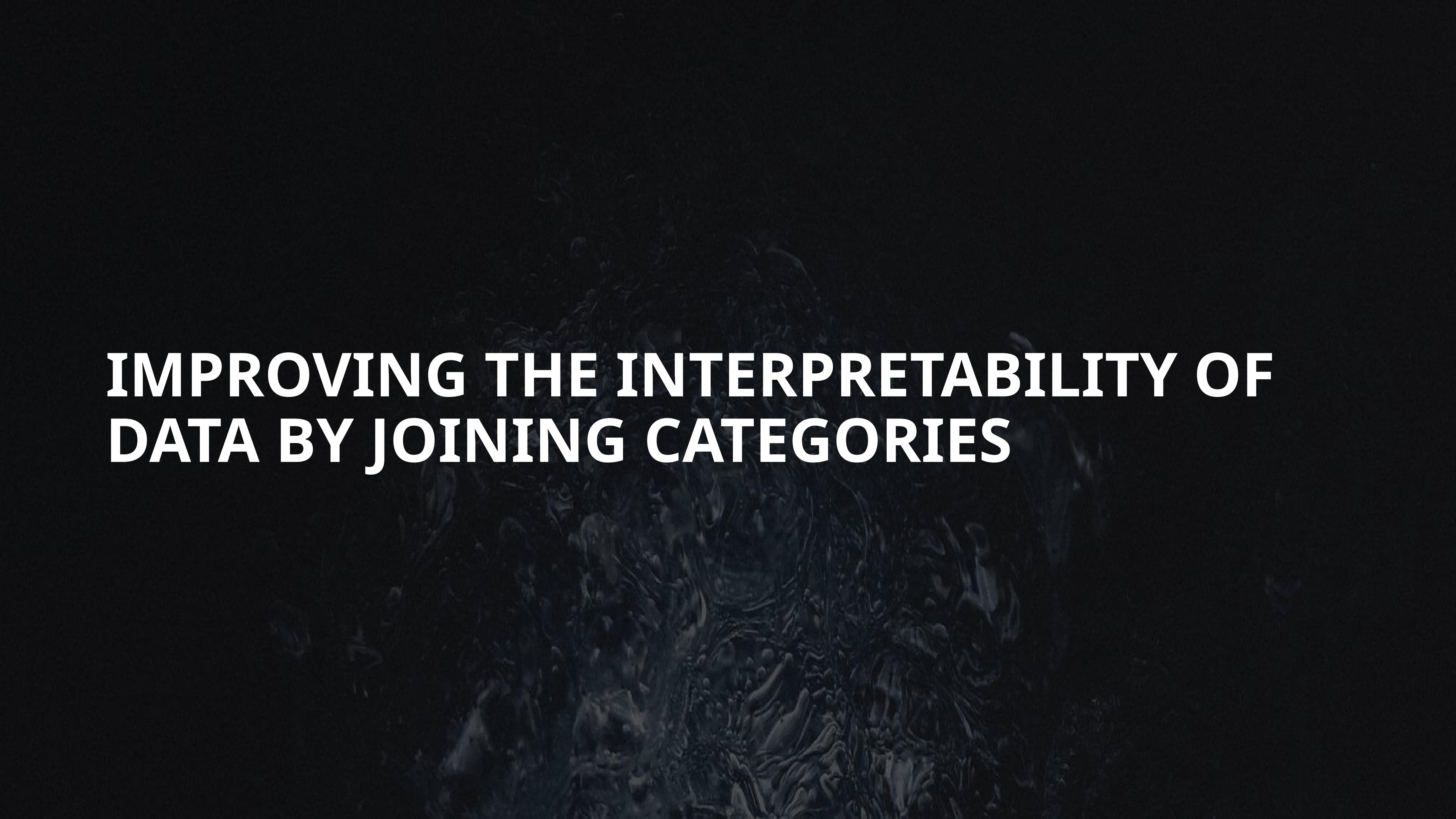
DATA CLEANING

For every modeling task, the data must have the following properties:

1. There must be no missing values.
2. All values must be numerical.
3. Categorical variables must be one-hot encoded (or dummy encoded).

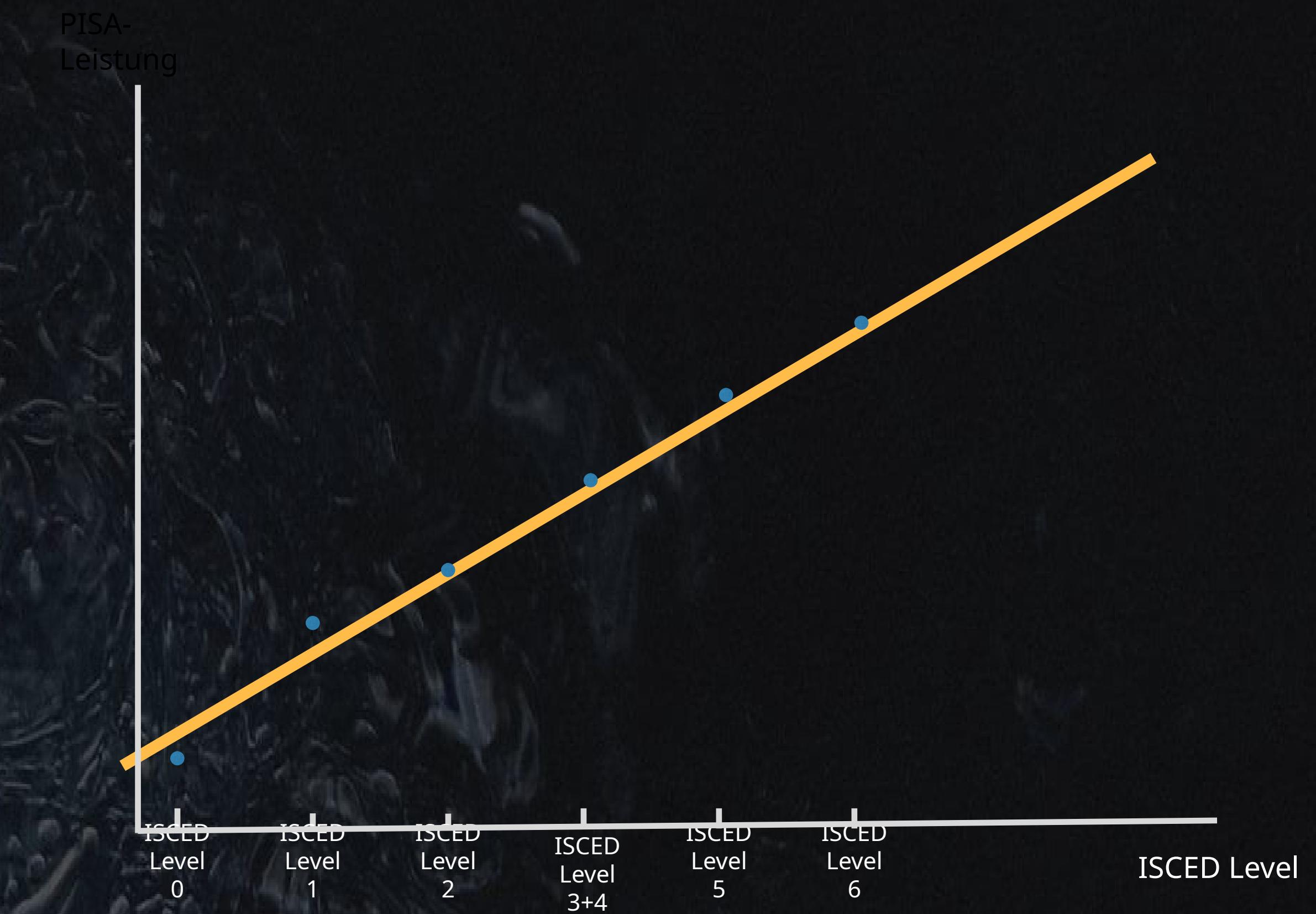
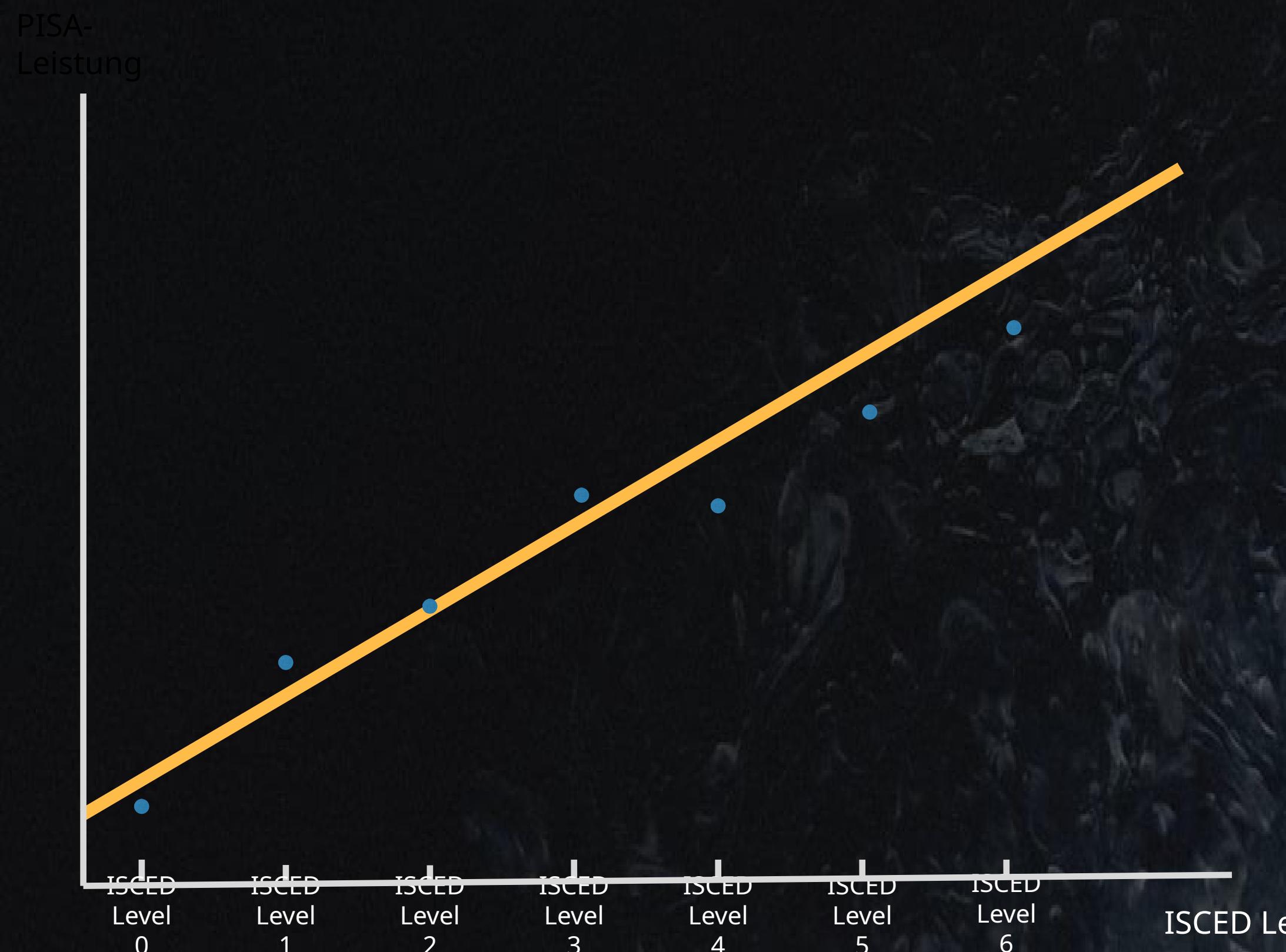
ONE-HOT ENCODING

id	color
1	red
2	blue
3	green
4	blue

The background of the slide features a dark, abstract pattern resembling liquid or smoke, with subtle variations in color and texture across the surface.

**IMPROVING THE INTERPRETABILITY
OF
DATA BY JOINING CATEGORIES**

IMPROVEMENT OF ORDINAL DATA ("BINNING")

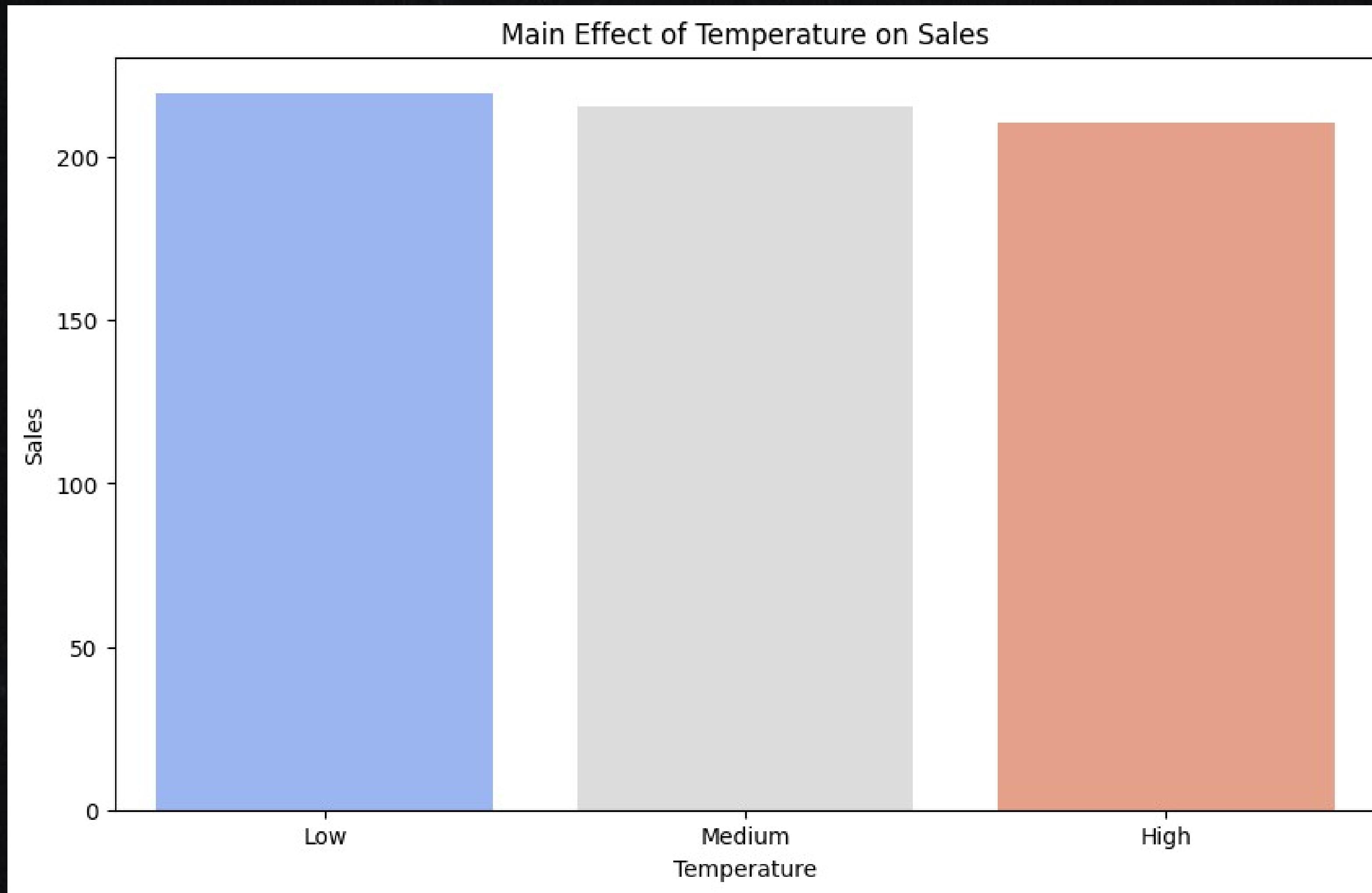


IMPROVING NOMINAL DATA

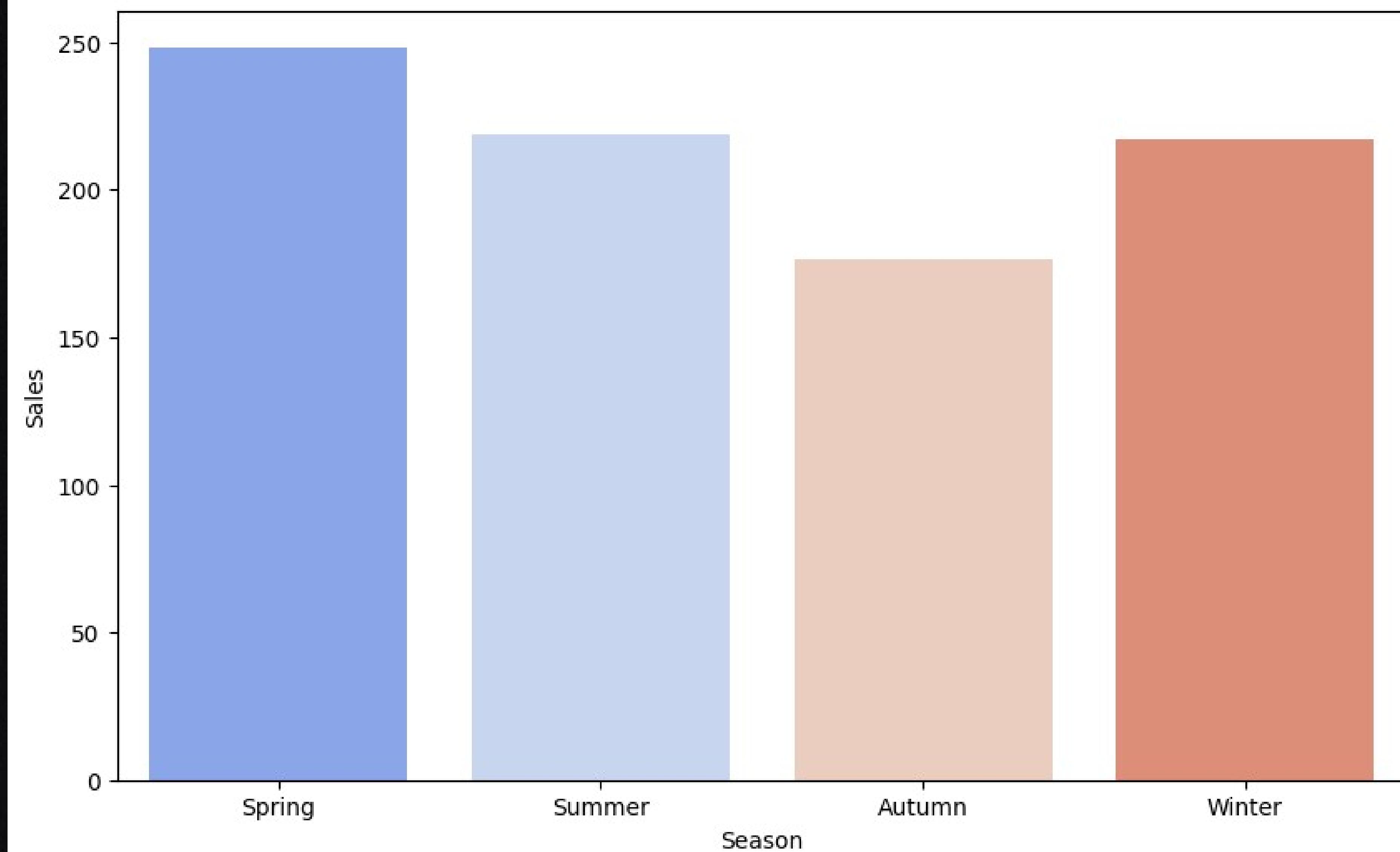
PISA-
Leistung

- Class sizes under 50 typically have very unstable mean values.
- A meaningful summarization of classes to a minimum size of approximately 50.
- Possibly use of a "residual category."

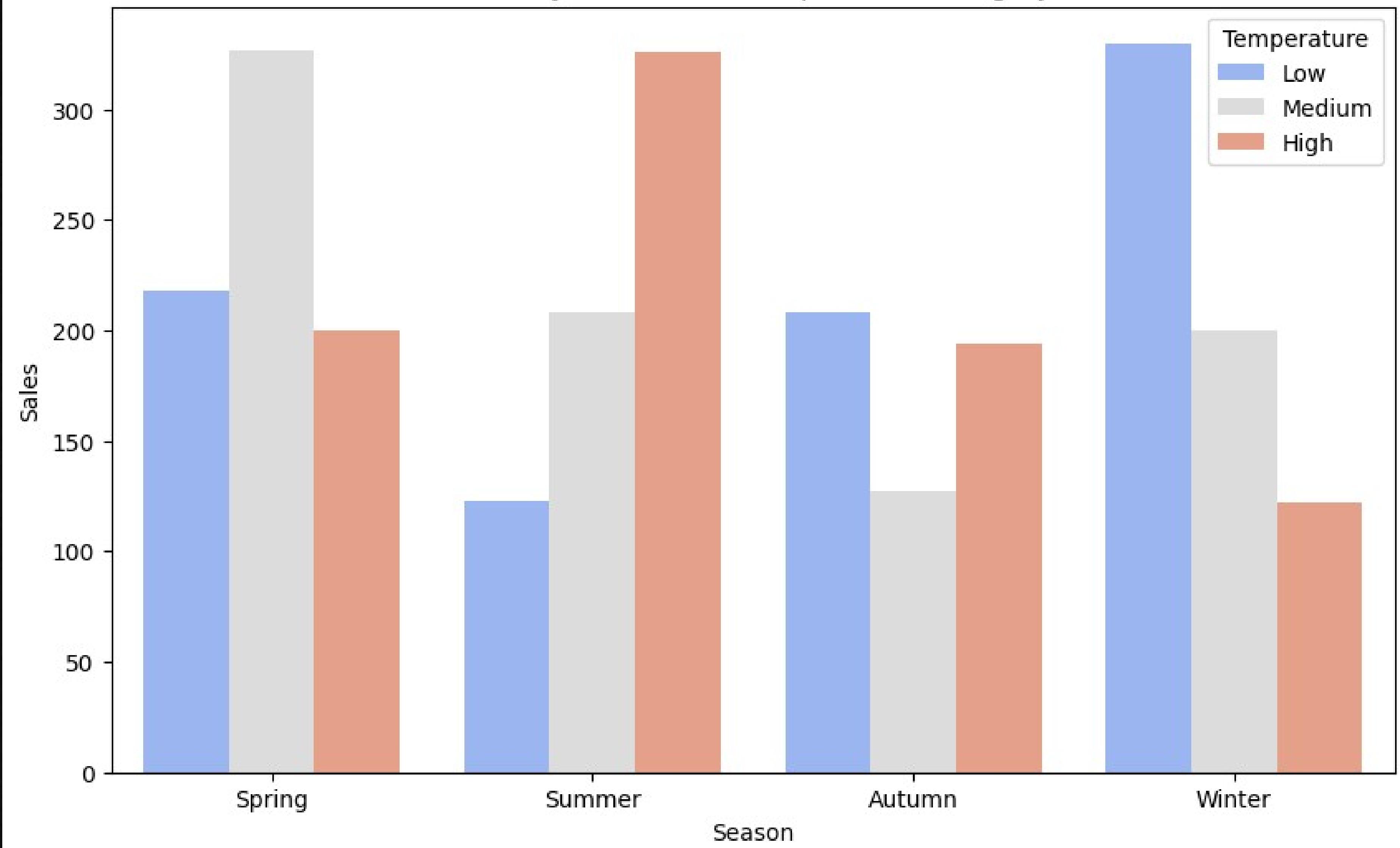
INTERACTION EFFECTS



Main Effect of Season on Sales



Sales by Season and Temperature Category



BREAKOUT

- What is the relationship between the daily temperature and sales volume?
- What are the ways to process the daily temperature as a variable in terms of 'Feature Engineering'?

INTRODUCTION TO TIME SERIES ANALYTICS

Moving average

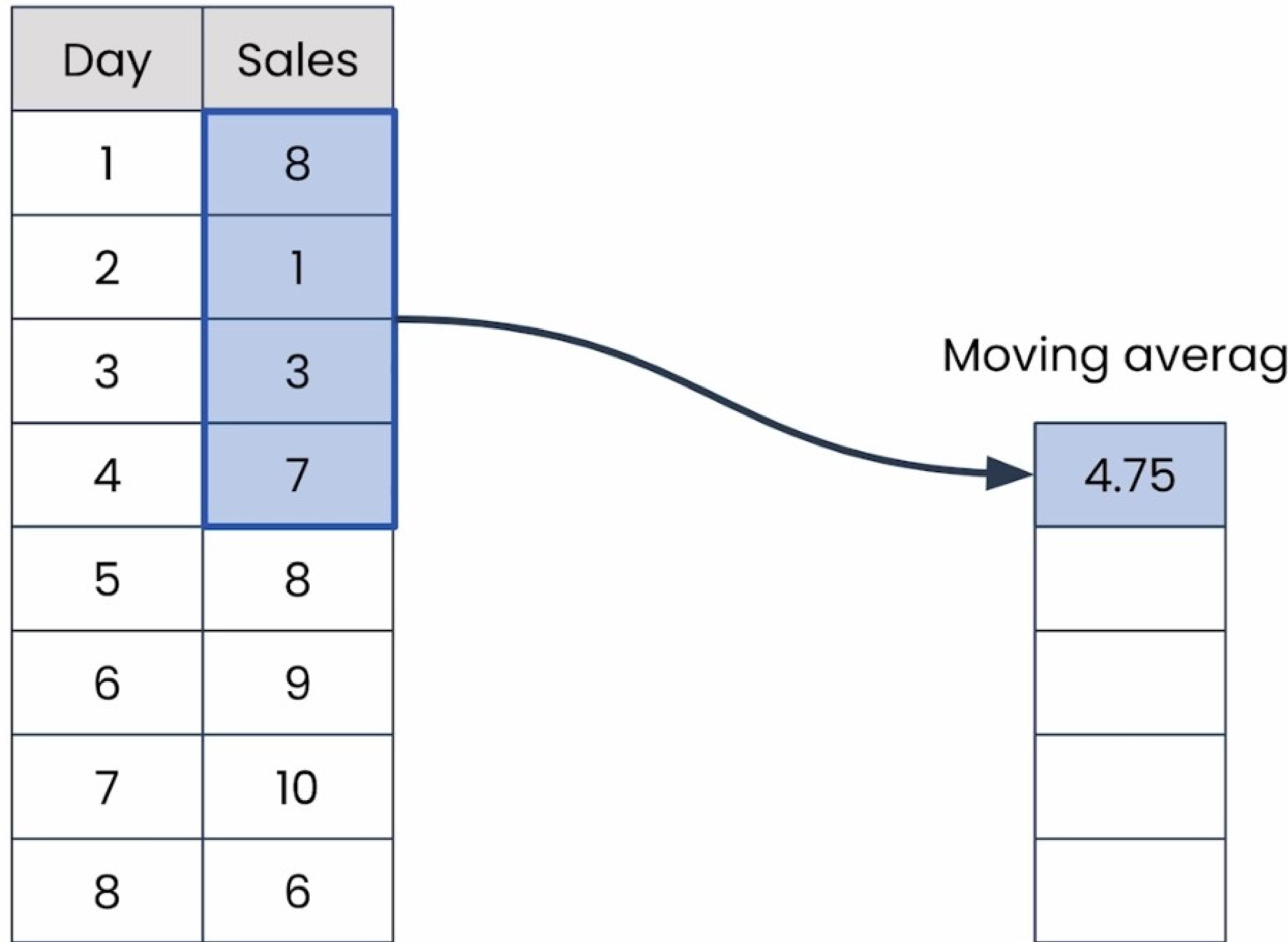
- Smooth out noisy behavior
- How to calculate a simple moving average:
 - Calculate average of **N** consecutive time periods
 - Calculates a series of values (**N-1** shorter than data)
 - Larger values of **N** are more stable

Day	Sales
1	8
2	1
3	3
4	7
5	8
6	9
7	10
8	6

Calculate average for values

Moving average

N = 4



Moving average

N = 4

Day	Sales
1	8
2	1
3	3
4	7
5	8
6	9
7	10
8	6

Moving average

4.75
4.75
6.75
8.50
8.25

Percent change

To standardize these differences:

- Convert from original units to percentages:

$$\frac{\text{Sales}_{\text{current}} - \text{Sales}_{\text{previous}}}{\text{Sales}_{\text{previous}}}$$

Day	Sales
1	8
2	1

Difference: -7 sales

$$\% \text{ change} = \frac{-7}{8}$$

Day	Sales
20	108
21	101

Difference: -7 sales

$$\% \text{ change} = \frac{-7}{108}$$

LEARNING RESOURCES

watch the following videos from the DeepLearning.AI course on Python for Data Analytics to get an idea about how to analyze time series data:

- [DateTimes](#) (5 minutes)
- [Using DateTimes as Indices](#) (3 minutes)
- [Moving Averages](#) (5 minutes)
- [Percent Change](#) (4 minutes)
- [Segmentation](#) (5 minutes)

(You need to create a free account with DeepLearning.AI.)

TASKS

- Meet with your team to discuss potential additional variables to be created for sales prediction (including whether there are any additional data sources you can use).
- Update the "Data Import and Preparation" directory in your team repository to include:
 - Additional downloaded or self-created data (e.g., holiday lists)
 - Code to merge all data into one dataset
 - Code to create new variables or prepare existing variables for prediction
- Get A Kaggle account and review the course competition under:
<https://www.kaggle.com/t/d832497c95d744de915ad9eb80ad9ed6>