

15.05.25

# Introduction to Data Science and Machine Learning

**VERSIONING WITH GIT (PART 2) AND  
CURRENT DEVELOPMENTS IN THE  
FIELD OF ML**

- **Addition of Teams**
- **Discussion of Tasks**
- **Introduction to Git - Part 2**
- **Additional Comments for the Data Preparation**
- **Current Developments in AI and Machine Learning**

# **ADDITION OF TEAMS**

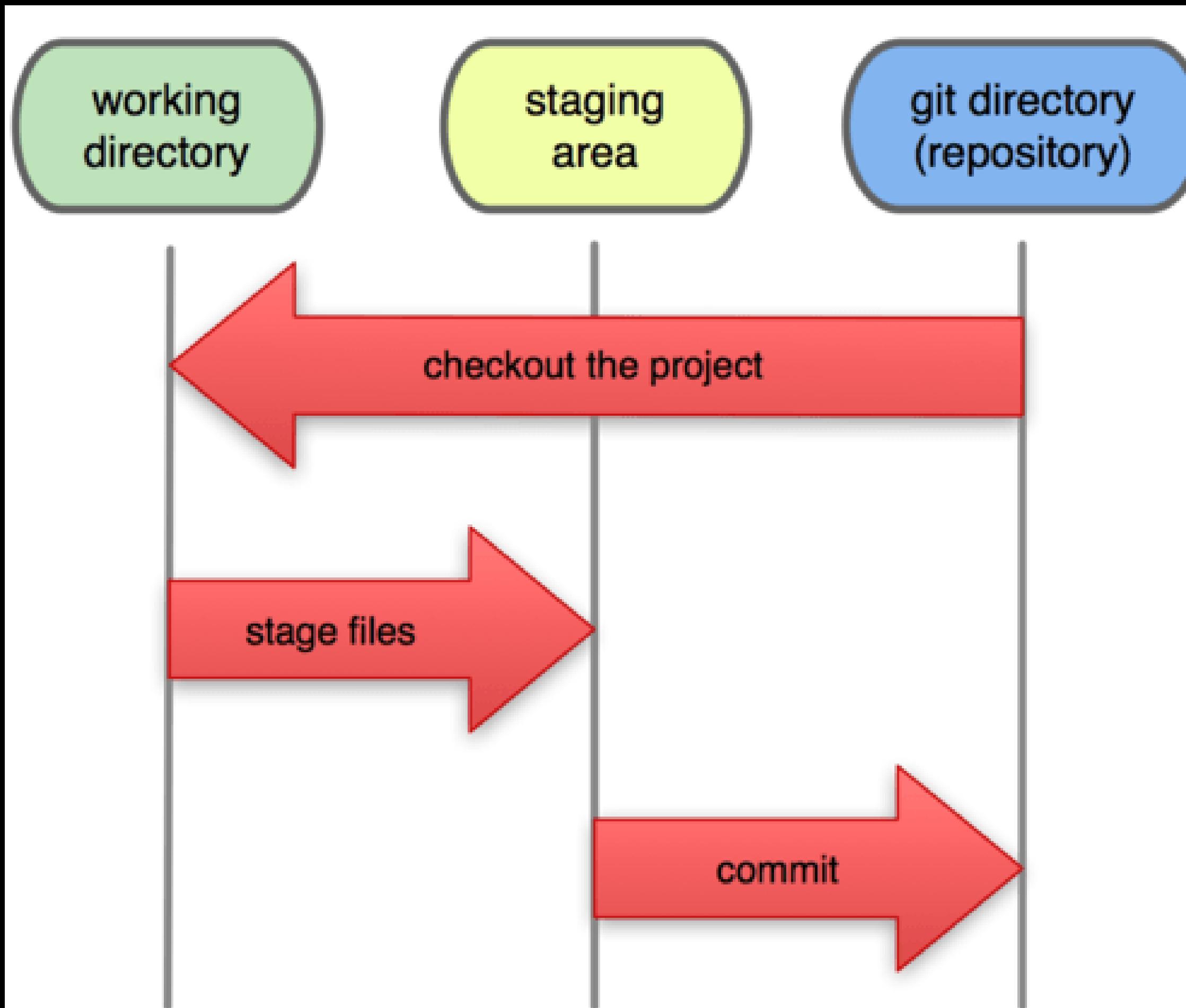
# BREAKOUT

- **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?

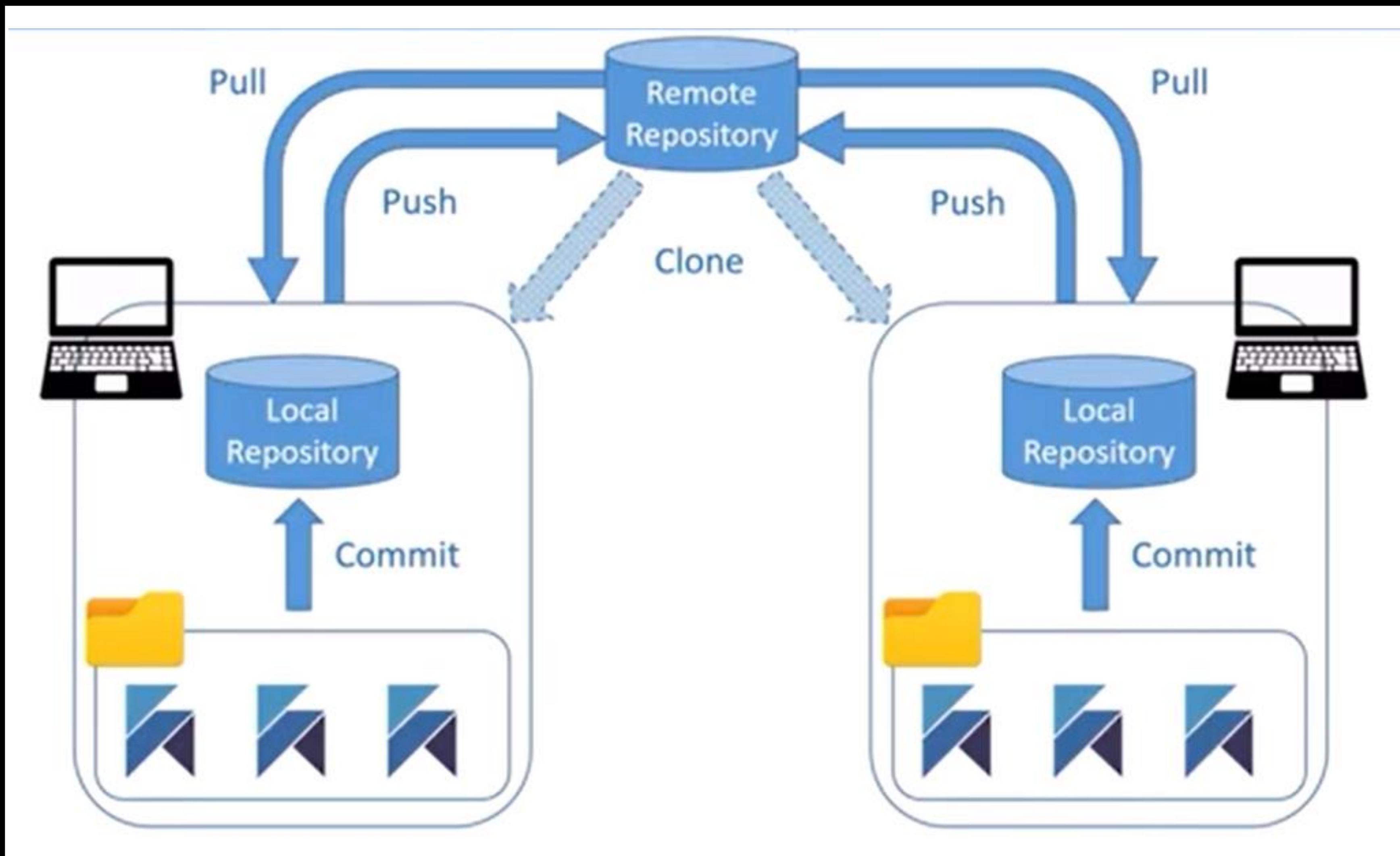
# VERSIONING WITH GIT



**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.

# LOCAL VS CODE

**Each commit is linked to a username and email. In GitHub Codespaces, these are automatically set.**

**For a local installation, you can define them as follows:**

```
git config --global user.name "your_username"  
git config --global user.email your_email@example.com
```

**To verify, you can use the following command:**

```
git config --list
```

# 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` and is a public template. The main interface includes a navigation bar with links for Code, Issues, Pull requests, Actions, Security, Insights, and Settings. Below the navigation is a search bar and a toolbar with options like Edit Pins, Watch, Fork, Star, and a prominent green "Use this template" button. A dropdown menu from the template button offers options to "Create a new repository" or "Open in a codespace". The repository's activity feed lists several commits from user `steffen74`, including updates to README.md and various subdirectories. The repository details sidebar on the right provides information such as the repository template for the course "Einführung in Data Science und maschinelles Lernen", statistics (2 stars, 4 watching, 1 fork), and links to Readme, Activity, Custom properties, and Report repository. The releases and packages sections indicate no releases or packages have been published.

opencampus-sh / repo-template-intro-to-data-science-and-ml

Type / to search

Code Issues Pull requests Actions Security Insights Settings

repo-template-intro-to-data-sci... Public template

Edit Pins Watch 4 Fork 1 Star 2 Use this template

Create a new repository Open in a codespace

main 1 Branch 0 Tags

Go to file

Code

steffen74 Update README.md c6965ea · 7 months ago 5 Commits

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

README

About

Repository Template für den Kurs Einführung in Data Science und maschinelles Lernen

Readme

Activity

Custom properties

2 stars

4 watching

1 fork

Report repository

Releases

No releases published

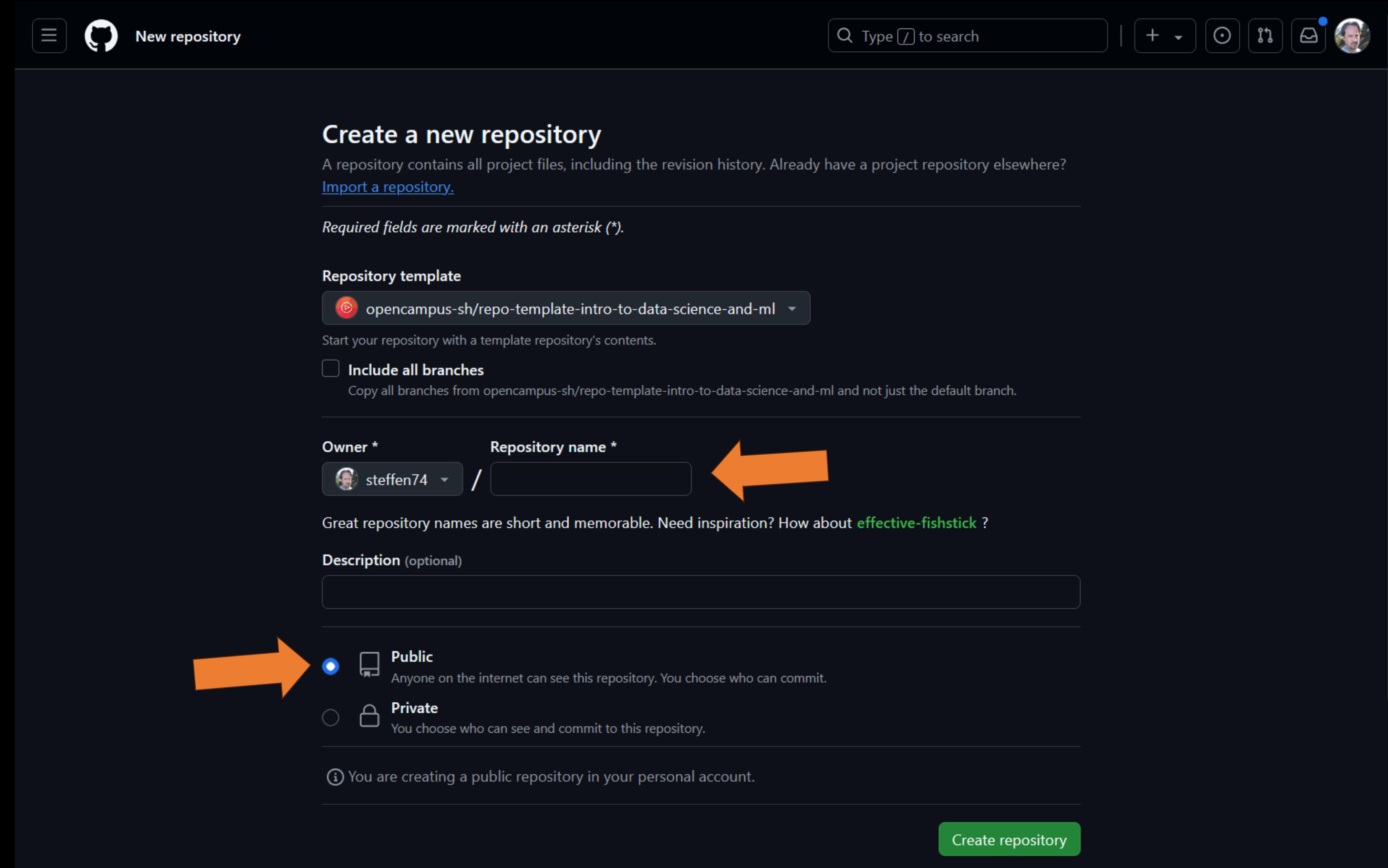
Create a new release

Packages

No packages published

Publish your first package

[https://github.com/new?template\\_name=repo-template-intro-to-data-science-and-ml&template\\_owner=opencampus-sh](https://github.com/new?template_name=repo-template-intro-to-data-science-and-ml&template_owner=opencampus-sh)





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](#)

## General

### Repository name

my\_repo

[Rename](#)

### Template repository

Template repositories let users generate new repositories with the same directory structure and files. [Learn more about template repositories.](#)

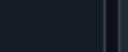
### Require contributors to sign off on web-based commits

Enabling this setting will require contributors to sign off on commits made through GitHub's web interface. Signing off is a way for contributors to affirm that their commit complies with the repository's terms, commonly the [Developer Certificate of Origin \(DCO\)](#). [Learn more about signing off on commits.](#)

## Default branch

The default branch is considered the "base" branch in your repository, against which all pull requests and code commits are automatically made, unless you specify a different branch.

main



## Social preview

Upload an image to customize your repository's social media preview.

Images should be at least 640×320px (1280×640px for best display).

[Download template](#)

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

Select a collaborator above

Manage

to this repository.

Manage access

You haven't invited any collaborators yet

Add people

The screenshot shows a GitHub repository interface with a dark theme. On the left, there's a sidebar with various settings like General, Access, Code and automation, Security, and Integrations. The 'Collaborators' tab is selected. A modal window is open over the page, titled 'Add a collaborator to my\_repo'. It contains a search bar with the placeholder 'Search by username, full name, or email' and a button labeled 'Find people'. Below the search bar is a green button that says 'Select a collaborator above'. At the bottom of the modal, there's another green button labeled 'Add people'. Two orange arrows are overlaid on the image: one pointing to the 'Find people' search bar and another pointing to the 'Add people' button.

# 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.**
- **If you still have time, each of you should also push the code for your descriptive statistics and visualizations**

# 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 PREPARATION

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

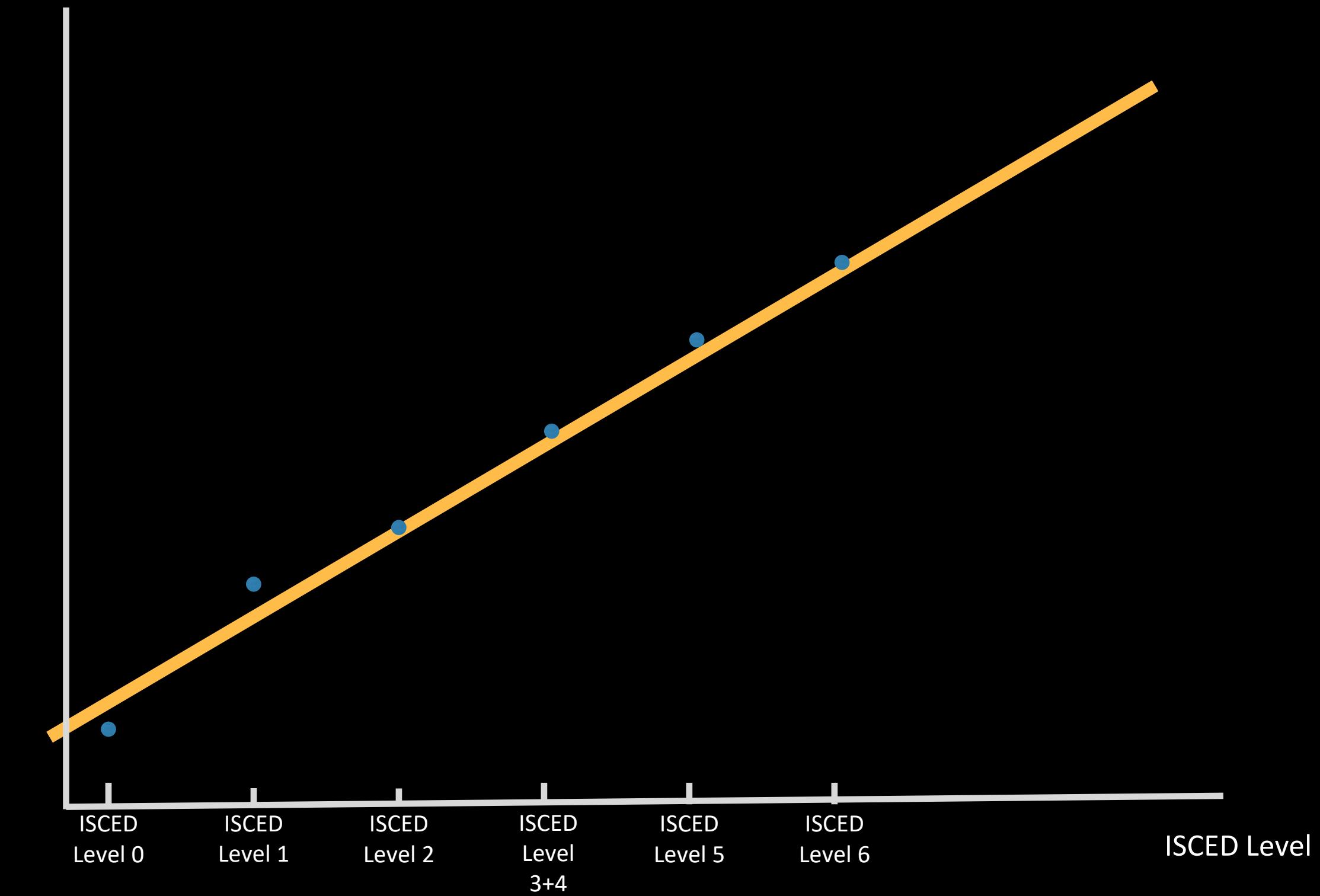
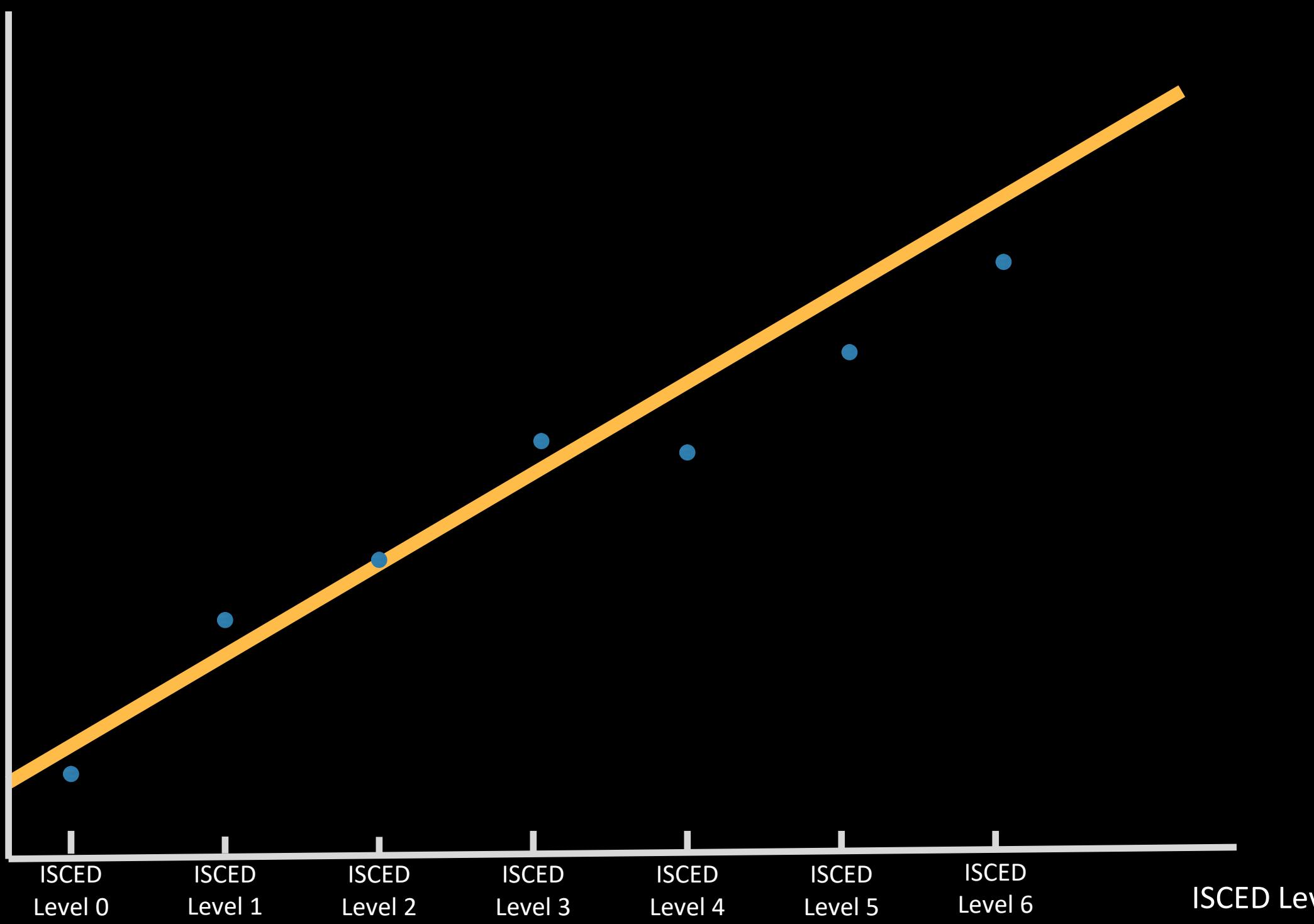
- 1. There must be no missing values.**
- 2. All values must be numerical.**
- 3. Categorical variables must be one-hot encoded (or dummy encoded).**

# ONE-HOT ENCODING

id	color
1	red
2	blue
3	green
4	blue

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

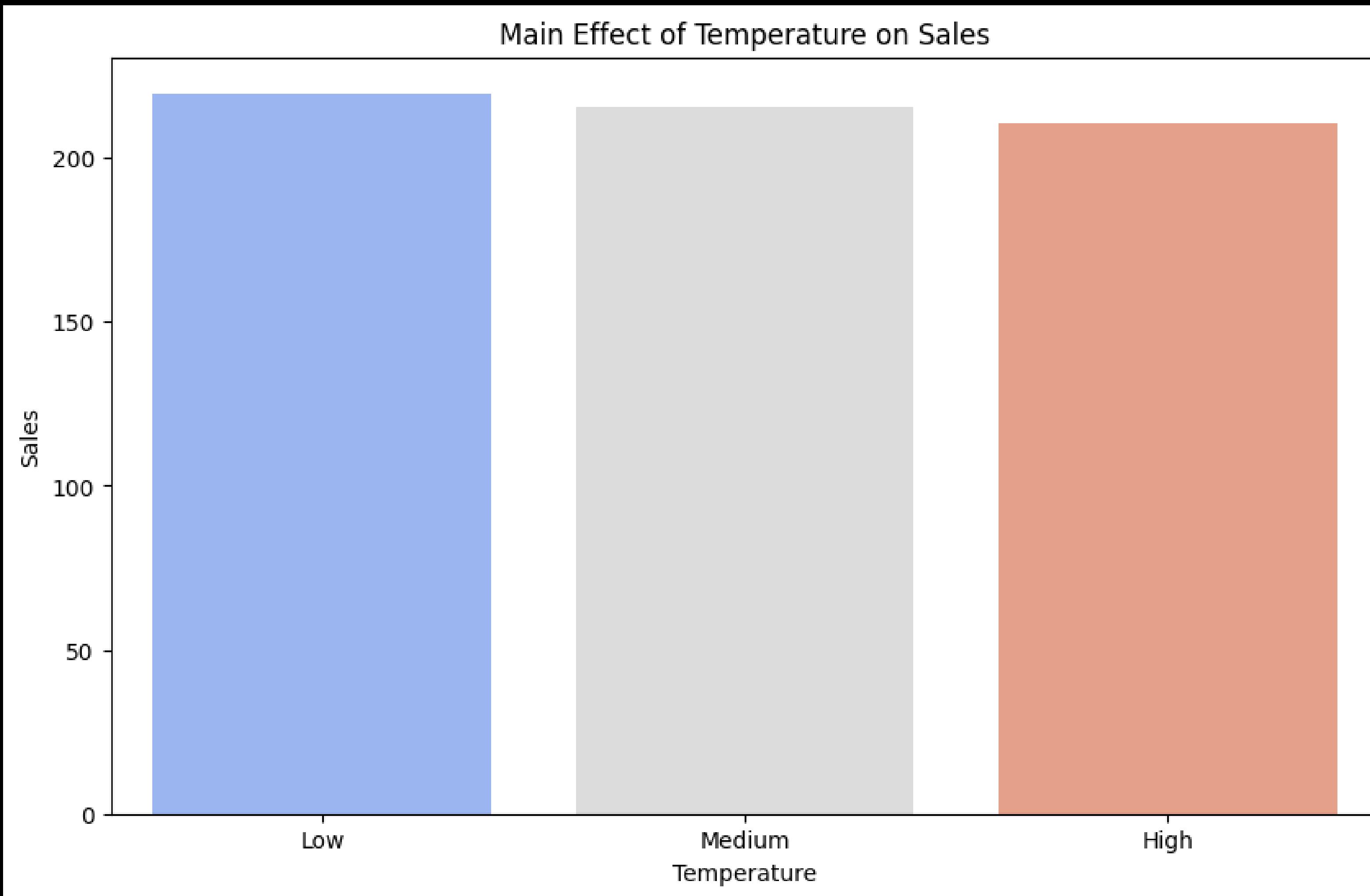
# IMPROVEMENT OF ORDINAL DATA ("BINNING")



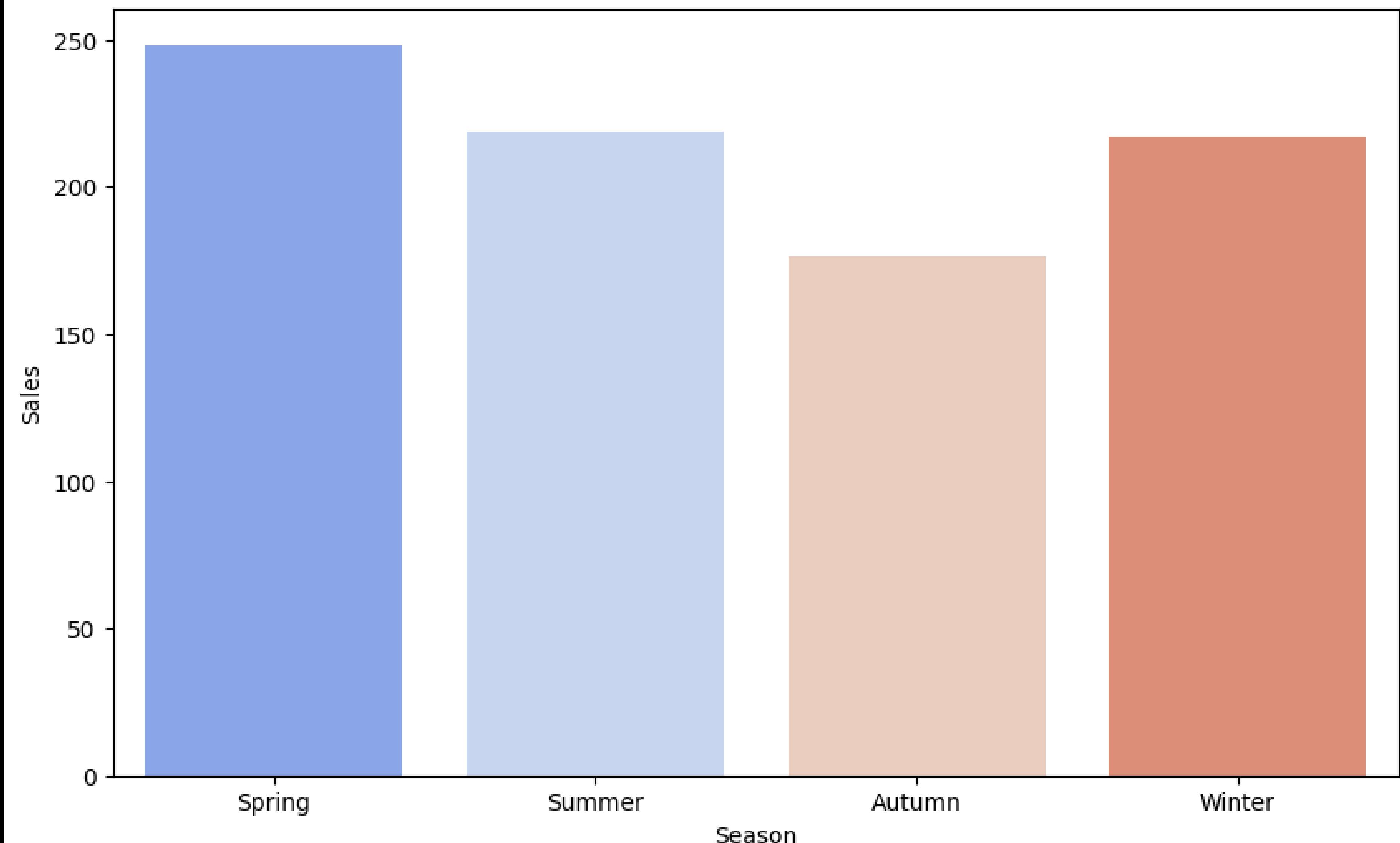
# IMPROVING NOMINAL DATA

- **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."**

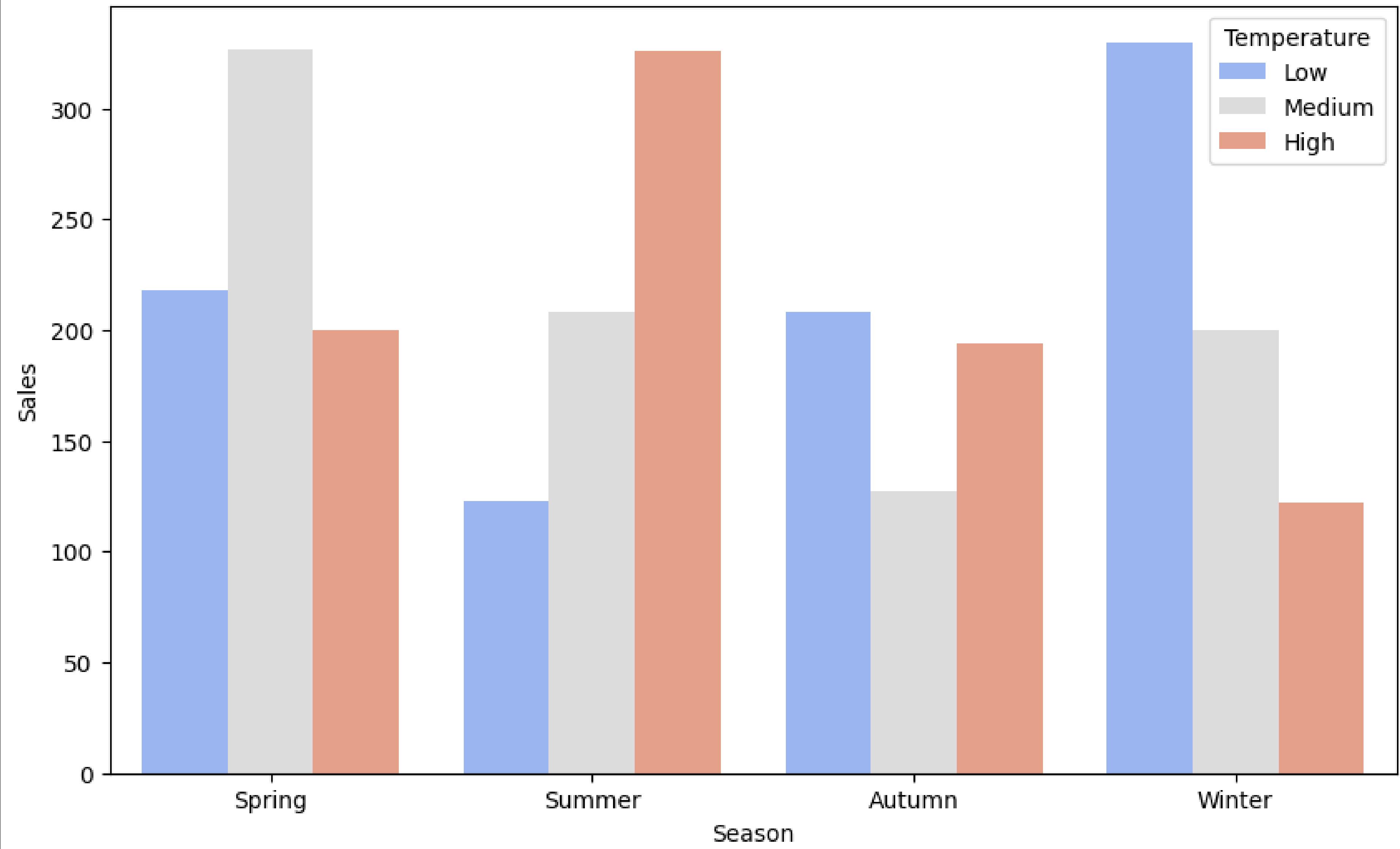
# INTERACTION EFFECTS



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'?**

# LERNMATERIAL

- Complete the sections on „Linear Regression“ and „Multiple Linear Regression“ from Module 4 of the course Python for Data Analytics by Deeplearning (~2 hours).

# **HOW TO REGISTER FOR THE COURSE FOR FREE**

[For Individuals](#)[For Businesses](#)[For Universities](#)[For Governments](#)[Explore](#)

What do you want to learn?

[Degrees](#)[Careers](#)[English](#)

S

[Home](#) > [Browse](#) > [Data Science](#) > [Data Analysis](#)

DeepLearning.AI

# Python for Data Analytics

This course is part of [DeepLearning.AI Data Analytics Professional Certificate](#)



Instructor: [Sean Barnes](#)

**Enroll for Free**  
Starts May 14

5,311 already enrolled

Try for Free: Enroll to start your 7-day full access free trial

**5 modules**

Gain insight into a topic and learn the fundamentals.

**Beginner level**

Recommended experience

**48 hours to complete**

3 weeks at 16 hours a week

**Flexible schedule**

Learn at your own pace

[About](#)[Outcomes](#)[Modules](#)[Recommendations](#)[Testimonials](#)



# Python for Data

This course is part of [DeepLearning.AI Data](#)



Instructor: [Sean Barnes](#)

**Enroll for Free**  
Starts May 14

5,311 already enrolled

## 5 modules

Gain insight into a topic and learn the fundamentals.

[Audit](#) this course

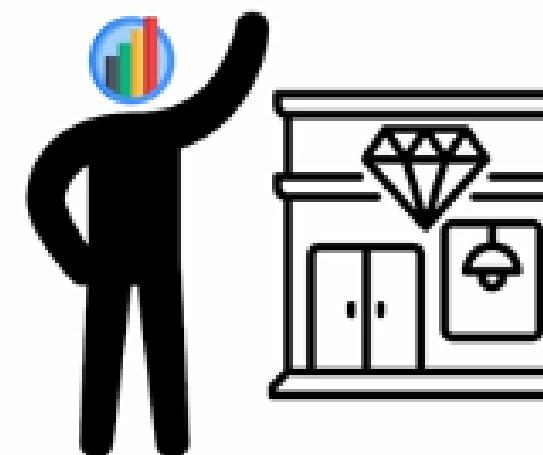


[Hide menu](#)**Introduction****Confidence intervals  
and hypothesis testing****Linear regression** **Video:** What is linear regression?  
6 min **Video:** Choosing an independent variable  
3 min **Video:** Training the model  
5 min **Video:** Interpreting the output of a regression model  
4 min **Video:** Prediction  
4 min **Practice Assignments:**  
Lesson 2 quiz  
10 min **Lab:** Practice Lab:  
London housing prices -  
Linear regression  
30 min

Python for Data Analytics &gt; Module 4 &gt; What is linear regression?

< Previous Next >

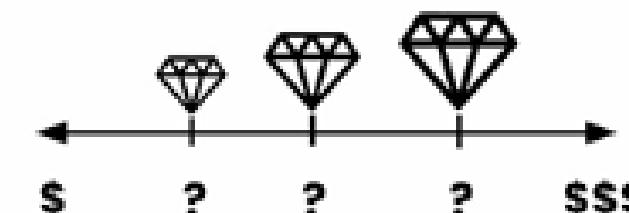
## Scenario



You  
Data Analyst

**Goal:** Pricing new diamonds acquired by the company **Task:** Predict market price for each diamond based on:

- Size
- Cut
- Clarity



Predict new prices for individual diamonds

Start with simple model using just one factor, then add more later

Sean Barnes

1x



Report an issue

## What is linear regression?

Save note

TranscriptNotesDownloads

Now that you're familiar with confidence intervals and hypothesis tests, you're ready to expand

Transcript language: English (US)

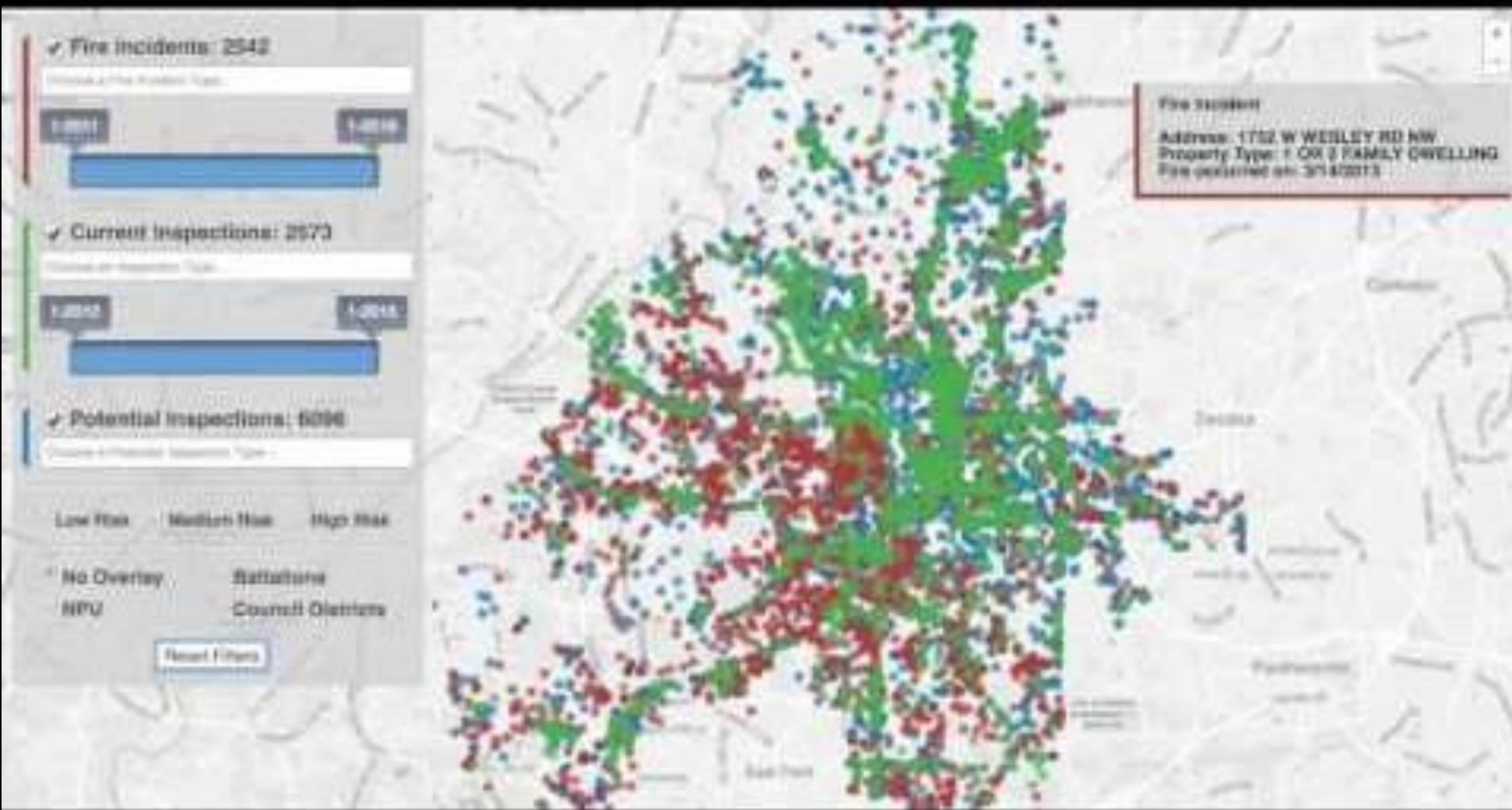


# TASKS

- **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**

# CURRENT AI APPLICATIONS

# RISK FORECASTING



# CHESS/ GAMING COMPUTER



AlphaGo – The Movie:

<https://www.youtube.com/watch?v=WXuK6gekU1Y>

# ROBOTICS





Unitree Go2  
New Upgrade

# **IMAGE AND VIDEO GENERATION**

# SORA

Explore

Search Ctrl K

Explore

Images

Videos

Top

Likes

Library

My media

Favorites

Uploads

Trash

New folder

manoftomorrow

Q 3.3K

amsoil64

Q 1.7K

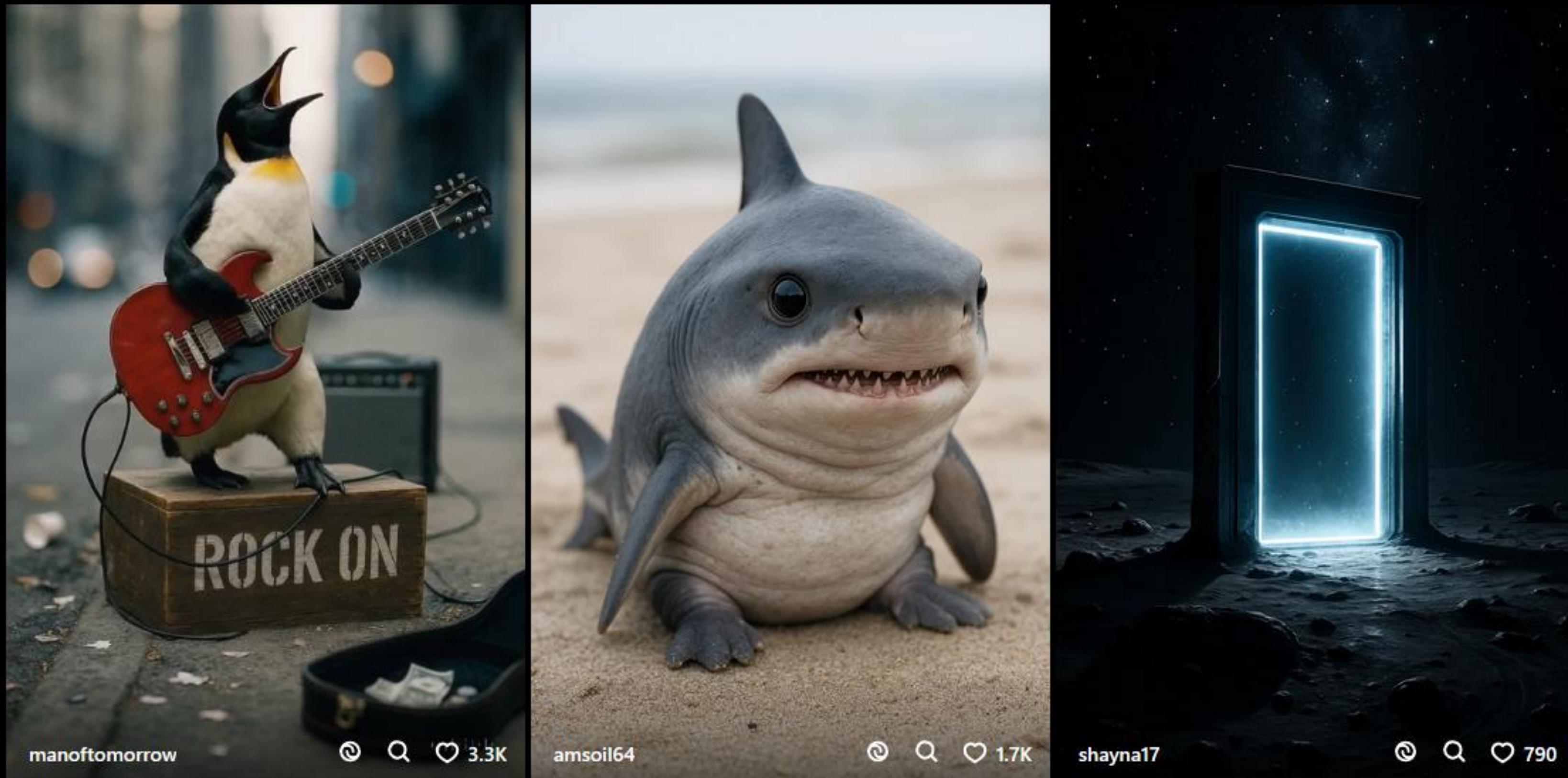
shayna17

Q 790

+ Describe your image...

Image 2:3 4v ?

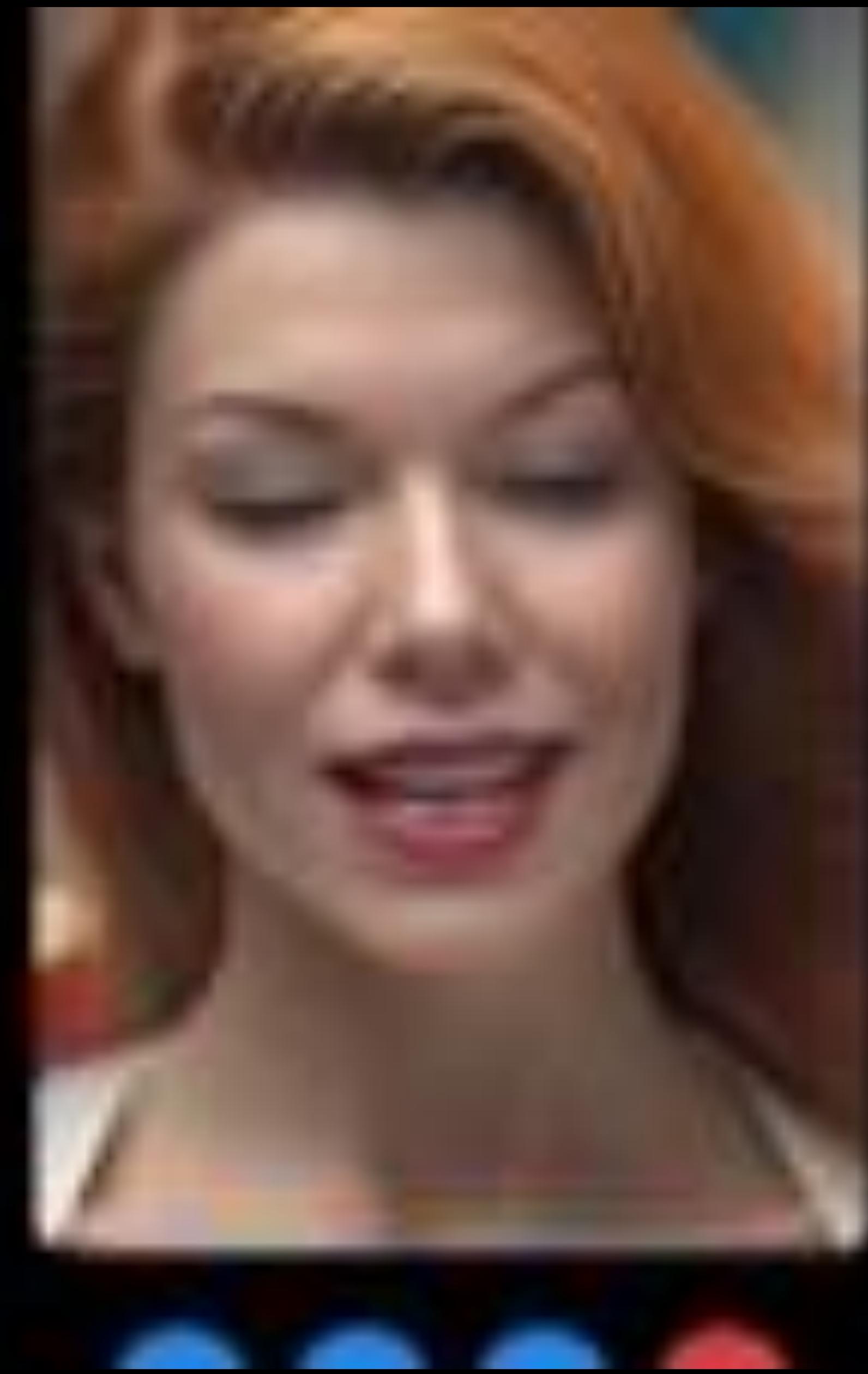
[https://sora.chatgpt.com/g/gen\\_01jv5t8sjfe0v80w468vbnxwfc](https://sora.chatgpt.com/g/gen_01jv5t8sjfe0v80w468vbnxwfc)



# HEY GEN

The screenshot displays the HeyGen video creation interface. On the left, a sidebar titled "Pick an Avatar" offers options like "My Avatar", "Instant Avatar" (selected), "Photo Avatar", and "Studio Avatar". Under "Instant Avatar", two demo options are shown: "Demo Finetune" and "Demo Finetune". Below this, the "Public Avatar" section shows "Instant Avatar", "Photo Avatar", and "Studio Avatar" again, with several 4K resolution options. At the bottom of the sidebar are "Pricing" and a purple circular icon.

The main workspace shows a woman with blonde hair in a dark blue shirt standing in front of a white couch and a large green plant. The HeyGen logo is visible in the bottom right corner of the video frame. The bottom of the screen features a timeline with a play button at 00:00:00:07, a preview window showing the woman, and a duration indicator of 7.7s. To the right of the timeline are controls for "Text Script" and "Audio Script", a message about starting with a script, and a character selection for "Jane - Serious" (Female, Youth, News, Narration...). It also includes sliders for "Speed (1.0x)" and "Pitch (0%)", a "Play Scripts" button, and a "Tips to improve pronunciations" link.

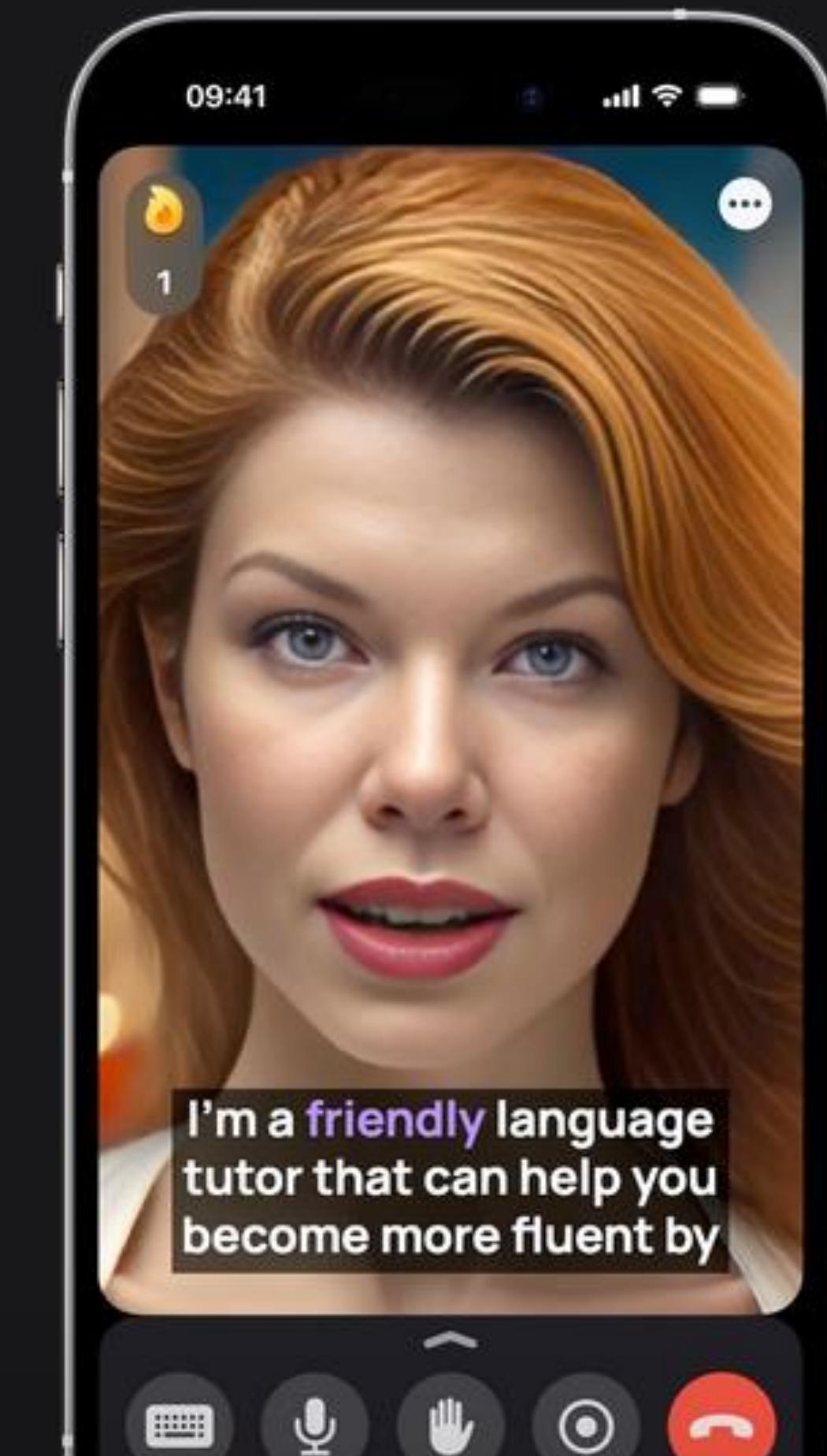


*Call Annie*

Sign in

# Speak English fluently

Video call our AI tutors and  
improve faster than ever.



Improve your  
pronunciation

Learn with a  
personal plan

Grow your  
vocabulary



Advay Pal [@advaypal]. (2023, April 26). what have we done? [Https://t.co/DVC0zL2Xgh](https://t.co/DVC0zL2Xgh) [Tweet]. Retrieved October 9, 2023, from Twitter website: <https://twitter.com/advaypal/status/1651355572980310016>

# NATURAL LANGUAGE PROCESSING

# GPT-3

## Explain Kants Categorical Imperative

↗ Share ⚙ Try again

The Categorical Imperative is an ethical theory proposed by the German philosopher Immanuel Kant. This was one of two main moral theories in his philosophy (the other being Deontology). His 'categorical imperative' proposes that an action can be considered moral if it could potentially become a universal law.

The Categorical Imperative is essentially the idea that the ends don't justify the means. Only actions which are consistent with a universal law (that can be applied consistently to all human beings) can be considered moral.

The Categorical Imperative also suggests that the motivations behind an action are irrelevant. For example, if I was to kill my enemy in cold blood, but did it out of compassion and love for humanity, this would not make it moral - as the act itself is contradicting a universal law.

However, it is important to note that the Categorical Imperative also says that an action can be considered moral if there are no other means of achieving a goal. For example, if I was in need of money for my own survival and I needed to steal it from someone else - without causing any harm.

Overall, I would say Kant's Categorical Imperative is a moral theory that puts great emphasis on the principle of 'doing unto others as you would have them do to you'.

It is also worth noting that Kant's Categorical Imperative disagrees with the Utilitarian theory of ethics, as it puts more emphasis on an action being consistent with a universal law rather than providing the greatest good to the greatest number.

# CHATBOTS AS AGENTS

# Wie kann ich dir helfen?

Stelle irgendeine Frage



Starte Reasoning für



Deep Research



Bild erstellen



Canvas

 Beim Schreiben und Programmieren  
zusammenarbeiten

Gemini



2.0 Flash ▾



◆ Gemini

2.0 Flash

Fast all-round help



2.5 Flash (preview)

Our next reasoning model built for speed

2.5 Pro (preview)

Reasoning, maths and code

Deep Research

Get in-depth research reports



Gemini Advanced

Upgrade



Ask Gemini



Deep Research

Canvas



Hello, Steffen

◆ Try Gemini Advanced



# CHATBOTS FOR PROGRAMMING

Edit Selection View Go Run Terminal Help main-page.html - coding-waterkant [SSH: ghost-blog-user] - Cursor

production.json main-page.html U ... New chat

@ Add context Plan, search, build anything

Agent Ctrl+I claude-3.7-sonnet

Ctrl+# for model menu

Auto-select Thinking

claude-3.5-sonnet claude-3.7-sonnet MAX claude-3.7-sonnet ✓ gemini-2.5-pro-preview-05-06 gemini-2.5-pro MAX gpt-4.1

Cursor helps you select the best premium model, based on performance and speed

```
> themes > edition > coding-waterkant-dev-files > main-page.html
<html lang="en">
<body>
    <!-- Program Slider Section -->
    <link rel="stylesheet" href="https://cdn.jsdelivr.net/npm/swiper@8.4.1/dist/css/swiper.min.css">
    <div class="content-grid-container">
        <h2>PROGRAM</h2>
        <div class="swiper program-slider">
            <div class="swiper-wrapper">
                <!-- Monday -->
                <div class="swiper-slide">
                    <div class="schedule-day">
                        <h3>Monday - July 1, 2024</h3>
                        <div class="schedule-item">
                            Patron Dirk Schrödter</b>
                        <div class="schedule-item">
                            Be Announced □ Stay Tuned
                        <div class="schedule-item">
                            Food and Music</span></div>
                    </div>
                </div>
                <!-- Tuesday -->
                <div class="swiper-slide">
                    <div class="schedule-day">
                        <h3>Tuesday - July 2, 2024</h3>
                        <div class="schedule-item">
                            div>
                        <div class="schedule-item">
                            span></div>
                        <div class="schedule-item">
                            Be Announced □ Exciting
                        <div class="schedule-item">
                            span></div>
                        <div class="schedule-item">
                            Be Announced □ Stay Curious
                        <div class="schedule-item">
                            span></div>
                    </div>
                </div>
            </div>
        </div>
    </div>
```

Past chats ▾

- Cheapest Options for Installing Ghost
- CSS Formatting Issues in Div Class
- Daily Program Slider Implementation

ghost-blog-user main\* 101 0t 0 0 △ 0 ⌂ 1 Ln 452, Col 1 (17832 selected) Spaces: 4 UTF-8 LF { HTML ⌂ Go Live Cursor Tab ✓ Pre

Merged

Adds New Tablegrid generic component #809

ericstumper merged 14 commits into `staging` from `develop` 19 hours ago



review-agent-prime bot commented 20 hours ago

...

## frontend-nx/apps/edu-hub/queries/generated /DegreeParticipantsWithDegreeEnrollments.ts

Consider adding comments to the newly added code to improve readability and maintainability. This will help other developers understand the purpose of the code and any important details they should be aware of.

[Create Issue](#)

```
// This is an example of how you might comment the new code
// URL to the file containing the user's attendance certificate (if he obtained one)
attendanceCertificateURL: string | null;
```



## frontend-nx/apps/edu-hub/queries/user.ts

Consider optimizing the query by only fetching the necessary fields. This will reduce the amount of data transferred and improve the performance of the application.

[Create Issue](#)

```
// This is an example of how you might optimize the query
User(where: { status: { _eq: ACTIVE } }) {
  id
  firstName
  lastName
}
```



Consider sanitizing the input to the query to prevent SQL injection attacks. This can be done by using parameterized queries or prepared statements.

[Create Issue](#)

```
// This is an example of how you might sanitize the input
query UserForSelection1($searchValue: String!) {
  User(where: { firstName: { _ilike: $searchValue } }) {
    id
    firstName
    lastName
  }
}
```



# MULTIMODAL MODELS



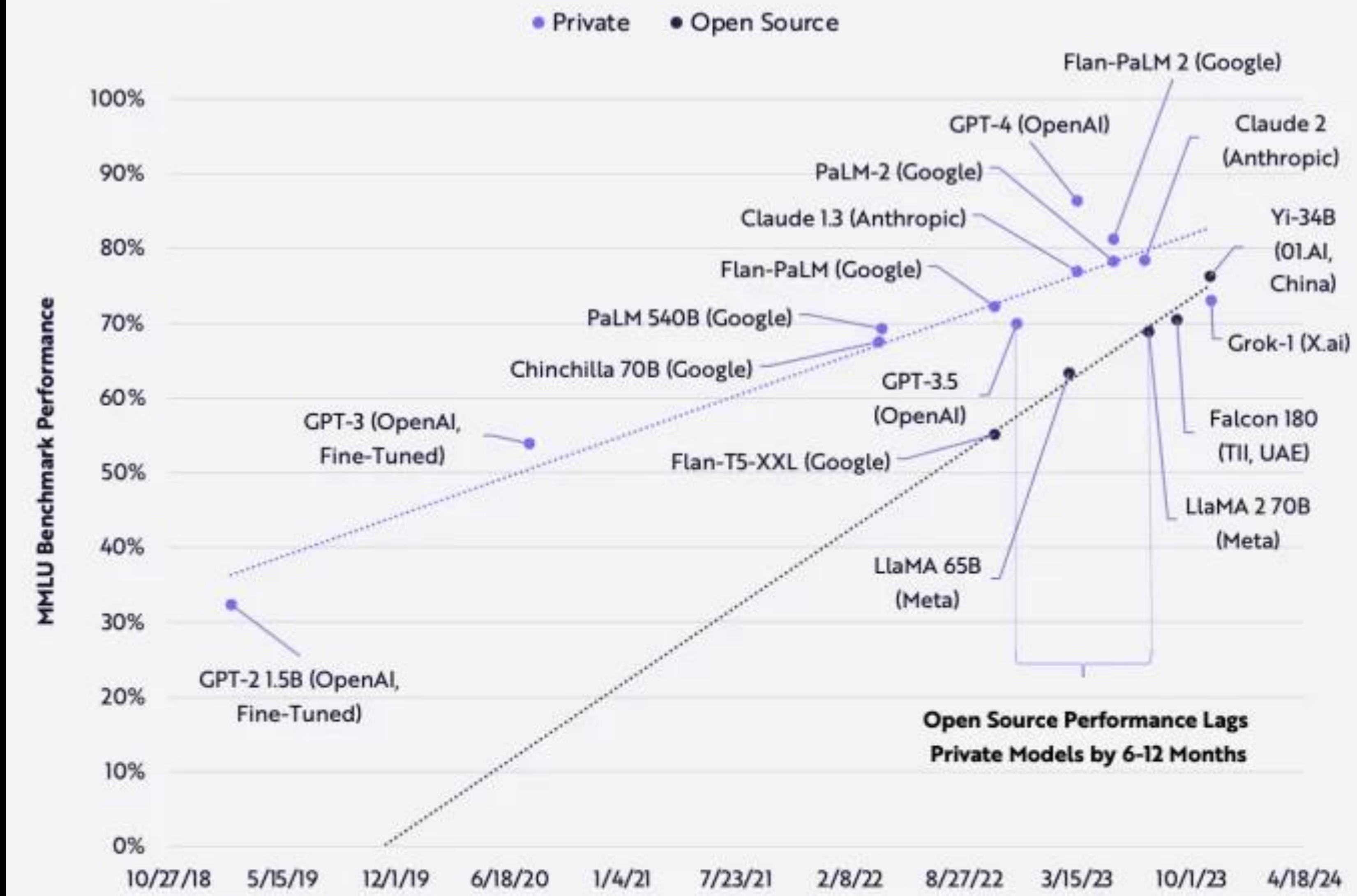
# CREATIVITY

**“A BOWL OF SOUP  
THAT LOOKS LIKE  
A MONSTER  
KNITTED OUT OF  
WOOL.”**

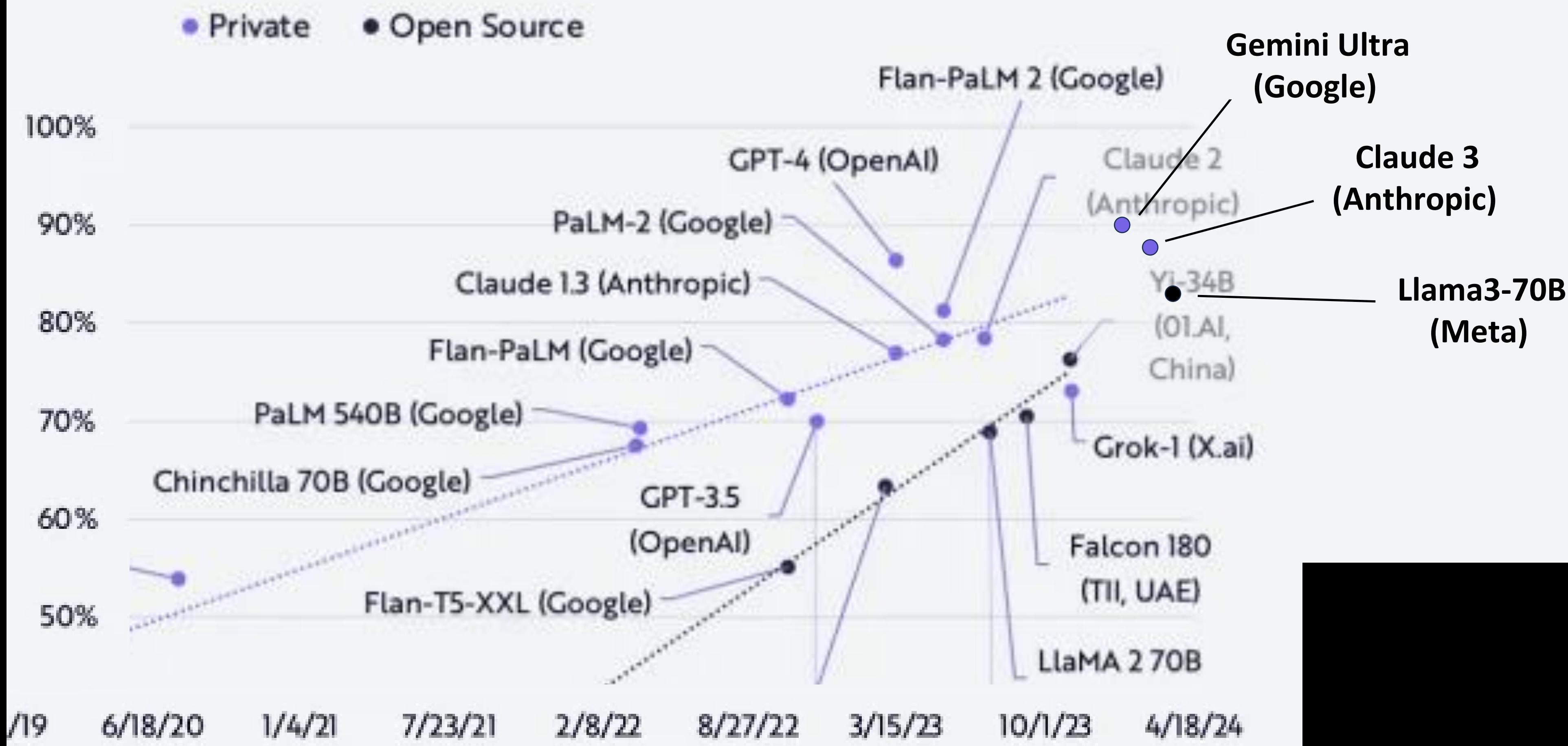
**“TEDDY BEARS  
MIXING SPARKLING  
CHEMICALS AS MAD  
SCIENTISTS IN A  
STEAMPUNK STYLE.”**

# OPEN SOURCE AI

## Open Source vs Private Models, 5-Shot MMLU Performance



## Private vs Private Models, 5-Shot MMLU Performance



# LLM LEADERBOARD

This site is currently in Beta testing. It may lack some features and be a bit buggy. Data loss may occur. [We'd love to hear your feedback!](#)

Overview    Text    WebDev    Vision    Text-to-Image    Search    Copilot

## Leaderboard Overview

See how leading models stack up across text, image, vision, and beyond. This page gives you a snapshot of each Arena, you can explore deeper insights in their dedicated tabs. Learn more about it [here](#).

[View Blog ↗](#)

**Text** ⌚ 2 days ago

Rank (UB) ↑	Model ↑	Score ↓	95% CI (±) ↓	Votes ↓	Organization ↑	License ↑
1	gemini-2.5-pro-preview-05-06	1446	+8/-9	4.500	Google	Proprietary
2	o3-2025-04-16	1413	+8/-7	6.689	OpenAI	Proprietary
2	chatgpt-4o-latest-20250326	1408	+6/-6	10.290	OpenAI	Proprietary
2	grok-3-preview-02-24	1403	+4/-4	14.843	xAI	Proprietary
3	gpt-4.5-preview-2025-02-27	1398	+4/-5	15.275	OpenAI	Proprietary
4	gemini-2.5-flash-preview-04-17	1394	+7/-7	5.959	Google	Proprietary



Search models, datasets, users...

Models Datasets Spaces Posts Docs Solutions Pricing Log In Sign Up

NEW Try Cohere Command R+ on HuggingChat



# The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

Text-to-Image Image-to-Text

Text-to-Video Visual Question Answering

Document Question Answering Graph Machine Learning

Computer Vision

Depth Estimation Image Classification

Object Detection Image Segmentation

Image-to-Image Unconditional Image Generation

Video Classification Zero-Shot Image Classification

Natural Language Processing

Text Classification Token Classification

Table Question Answering Question Answering

Zero-Shot Classification Translation

Summarization Conversational

Text Generation Text2Text Generation

Sentence Similarity

Audio

Text-to-Speech Automatic Speech Recognition

Audio-to-Audio Audio Classification

Voice Activity Detection

Tabular

Tabular Classification Tabular Regression

Reinforcement Learning

Reinforcement Learning Robotics

Models 469,541

Filter by name

meta-llama/Llama-2-70b

Text Generation Updated 4 days ago 25.2k 64

stabilityai/stable-diffusion-xl-base-0.9

Updated 6 days ago 2.01k 393

openchat/openchat

Text Generation Updated 2 days ago 1.3k 136

llyasviel/ControlNet-v1-1

Updated Apr 26 1.87k

cerspense/zeroscope\_v2\_XL

Updated 3 days ago 2.66k 334

meta-llama/Llama-2-13b

Text Generation Updated 4 days ago 328 64

tiiuae/falcon-40b-instruct

Text Generation Updated 27 days ago 288k 899

WizardLM/WizardCoder-15B-V1.0

Text Generation Updated 3 days ago 12.5k 332

CompVis/stable-diffusion-v1-4

Text-to-Image Updated about 17 hours ago 448k 5.72k

stabilityai/stable-diffusion-2-1

Text-to-Image Updated about 17 hours ago 782k 2.81k

Salesforce/xgen-7b-8k-inst

Text Generation Updated 4 days ago 6.18k 57

Trending on 😊 this week

Models

Spaces

Datasets

**Hugging Face**

Search models, datasets, users...

Models

Datasets

Spaces

Docs

Solutions

Pricing



Log In

Sign Up

**openai/whisper-large-v2** like 487

Automatic Speech Recognition

PyTorch

TensorFlow

JAX

Transformers

99 languages

whisper

audio

hf-asr-leaderboard

arxiv:2212.04356

License: apache-2.0

**Model card**

Files and versions

**Community** 42

Train

Deploy

Use in Transformers

## Whisper

Whisper is a pre-trained model for automatic speech recognition (ASR) and speech translation. Trained on 680k hours of labelled data, Whisper models demonstrate a strong ability to generalise to many datasets and domains **without** the need for fine-tuning.

Whisper was proposed in the paper [Robust Speech Recognition via Large-Scale Weak Supervision](#) by Alec Radford et al. from OpenAI. The original code repository can be found [here](#).

Compared to the Whisper large model, the large-v2 model is trained for 2.5x more epochs with added regularization for improved performance.

**Disclaimer:** Content for this model card has partly been written by the Hugging Face team, and parts of it were copied and pasted from the original model card.

[Edit model card](#)Downloads last month  
70,522

### Hosted inference API

Automatic Speech Recognition

Examples

[Browse for file](#) or [Record from browser](#) or[Realtime speech recognition](#)

Librispeech sample 1

[Compute](#)

This model could not be loaded by the inference API.

Model openai/whisper-large-v2 does not exist

&lt;/&gt; JSON Output

**Model details**