

Data Cleaning Workshop

We will start at 7:10PM!

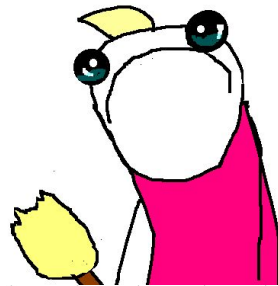


**DATA SCIENCE
COLLABORATIVE**

What is Data Cleaning?

Data that is collected will not always be easy for us to work with! We need to “clean” the data (among other things) to make it easy for us to visualize and analyze the data.

clean all the data?



Understanding the Data

The Pandas library gives us a very useful tool to see what parts of our data need to be cleaned!

```
netflix_data.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   show_id         8790 non-null   object
1   type            8790 non-null   object
2   title          8790 non-null   object
3   director       8790 non-null   object
4   country        8790 non-null   object
5   date_added     8790 non-null   object
6   release_year   8790 non-null   int64
7   rating         8790 non-null   object
8   duration       8790 non-null   object
9   listed_in     8790 non-null   object
dtypes: int64(1), object(9)
memory usage: 686.8+ KB
```

We can see missing data, and object types here

Understanding the Data

Recall that we can also take a quick look at the structure of our data

```
netflix_data.head()
```



	show_id	type	title	director	country	date_added
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021

We should always understand what our data looks like before trying to play with it!

Missing Data

Missing values can take form in many different ways

NaN

None

ERROR



A lot easier to handle!

Missing Data

When dealing with NONE or NaN, Pandas gives us simple commands to deal with these missing values:

Dropping rows where missing values are present:

```
netflix_data.dropna(inplace=True)
```

Replacing missing values in a column with mean/median:

```
netflix_data["release_year"].fillna(netflix_data["release_year"].mean())
```

Column we
want to fix

Function

Recall how to find the
mean of a column

Missing Data

Missing values can take form in many different ways

NaN

None

ERROR

Now let's try these!

Missing Data

First, let us identify these kinds of missing values:

```
pd.set_option('display.max_rows', None)
netflix_data["title"].value_counts()
```

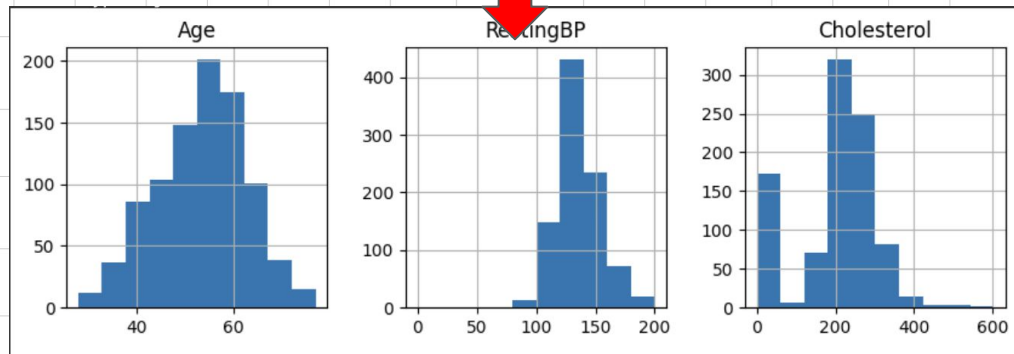


	count
title	
9-Feb	2
15-Aug	2
22-Jul	2
Dick Johnson Is Dead	1

Missing Data

For numerical variables, we can use a quick function to observe odd behavior in our variables

```
df.hist(figsize=(10, 10))
```



Missing Data

NOTE: This can be a tedious process if there are many differing missing value types

```
df["Cholesterol"] = df["Cholesterol"].replace(0, df["Cholesterol"].mean())
```

Column we
want to fix

Function

What value
we want to
replace

Column mean

In English:

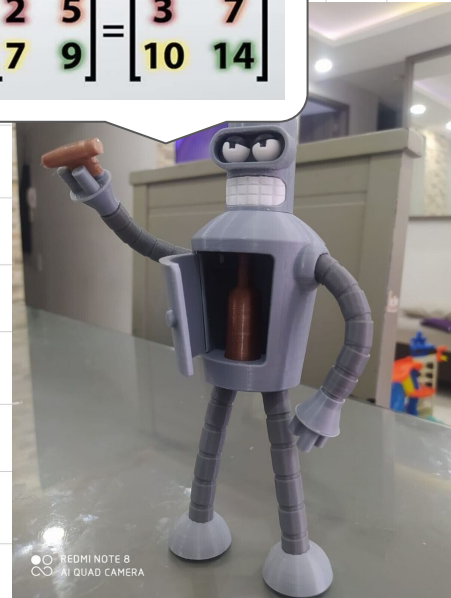
“In the **Cholesterol** column,
replace **0** with the **mean**”

Handling Categorical Data

Office Hours	
0	Yes
1	No
2	Yes
3	Yes
4	No
5	No
6	Yes
7	No
8	Yes
9	Yes

$$\begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} + \begin{bmatrix} 2 & 5 \\ 7 & 9 \end{bmatrix} = \begin{bmatrix} 3 & 7 \\ 10 & 14 \end{bmatrix}$$

A computer cannot distinguish any real meaning between “Yes” and “No”! We need to turn this into something the computer can understand



© NEDMI NOTE &
AI QUAD CAMERA

Handling Categorical Data

Office Hours	
0	Yes
1	No
2	Yes
3	Yes
4	No
5	No
6	Yes
7	No
8	Yes
9	Yes

```
df['Office Hours'] = np.where(df['Office Hours'] == "Yes", 1, 0)
```

Where value in "Office Hours" is equal to "Yes":
Replace with 1
Otherwise replace with 0

This ALWAYS works when there are only 2 different values

In other scenarios, we would handle this differently

Office Hours	
	1
	0
	1
	1
	0
	0
	1
	0
	1
	1

Handling Duplicate Observations

```
df.duplicated().sum()
```



```
1
```

We can remove duplicate observations using a simple function:

```
df.drop_duplicates(inplace=True)
```

Removing Variables

	show_id	type	title	director	country	date_added
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021

There is a simple function for this as well!

```
netflix_data.drop(columns={"show_id", "director"}, inplace=True)
```

If we have a column that numbers each observation, we will always drop it



Practice

<https://tinyurl.com/data-cleaning-2025>

Make a copy of the worksheet!



Announcements

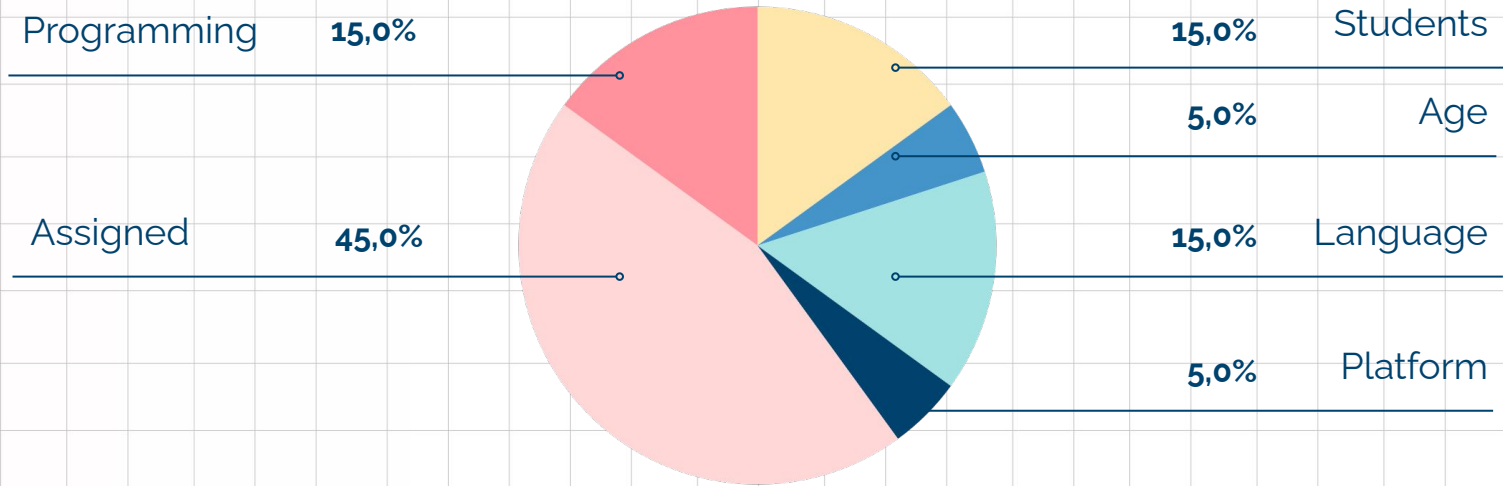


Apply to be an Intern!



Important Links!

Task Analysis



To modify this graph, click on it, follow the link, change the data and paste the new graph here

Validation of Tutoring Models

Advisor information

Mercury is the closest planet to the Sun

Data of the courses

Venus has a beautiful name, but it's hot

Tasks by role

Despite being red, Mars is a cold place

Administrative tasks

It's the biggest planet in the Solar System

Motivational tasks

Saturn is the ringed one and a gas giant

Organizational tasks

Neptune is the farthest planet from the Sun

You Can Use a Table

Mercury & Mars

Earth

- The third from the Sun
- It harbors life
- It's where we live on

Value

- 110
- 100
- 333

Value

- 029
- 200
- 080

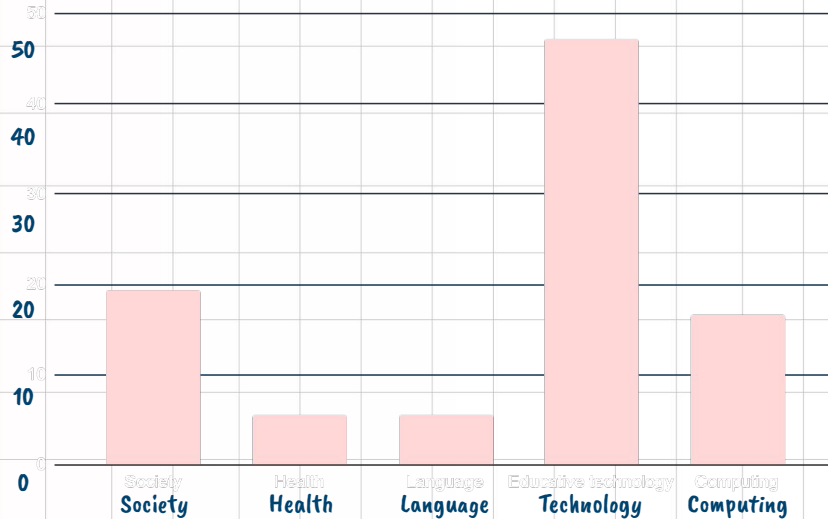
Saturn

- It has several rings
- It's a gas giant

- 018
- 382

- 003
- 033

Online Teaching



Mars

Mars is the fourth planet from the Sun and the second-smallest planet in the Solar System, being only larger than Mercury.

To modify this graph, click on it, follow the link, change the data and paste the new graph here

Greater Use of Distance Education

Canada

Planet Venus has a beautiful name

20%

Chile

Despite being red, Mars is a cold place

15%

South Africa

It's the closest planet to the Sun

10%

35%

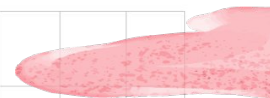
Russia

It's a gas giant and the biggest planet

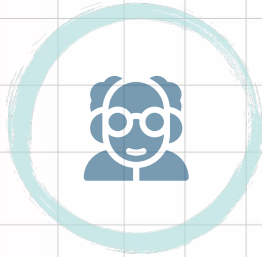
Australia

It's composed of hydrogen and helium

20%



A Timeline Always Works Well



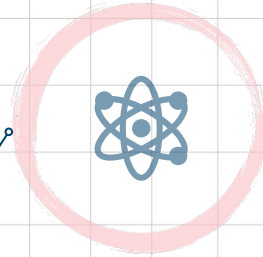
1895

Distance learning was first invented



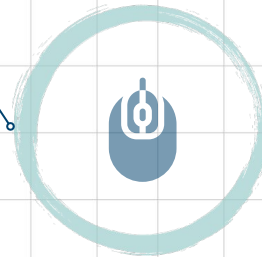
1982

Computer assisted learning center



1989

Internet becomes a household name



2006

3.5 million students using E-learning



19.7 M

Students enrolled in courses at degree-granting
postsecondary institutions in 2017




3,2M (19,5%)

Students took at least one but
not all of their courses online



2,2M (13,3%)

Students enrolled in exclusively
distance education



**274,211
(9,1%)**

Students took more than half of
their courses online



Course Start-development Tasks

01
Venus

It's the second planet from the Sun

02
Mars

Despite being red, it's a very cold place

03
Jupiter

It's the biggest planet in the Solar System

04
Mercury

Mercury is the closest planet to the Sun

Our Team



John Doe

You can replace the star
on the screen with a
picture of this person



Helena Patterson

You can replace the star
on the screen with a
picture of this person



Anton Smith

You can replace the star
on the screen with a
picture of this person



Conclusions

1

It's terribly hot—even hotter than Mercury—and its atmosphere is extremely poisonous

2

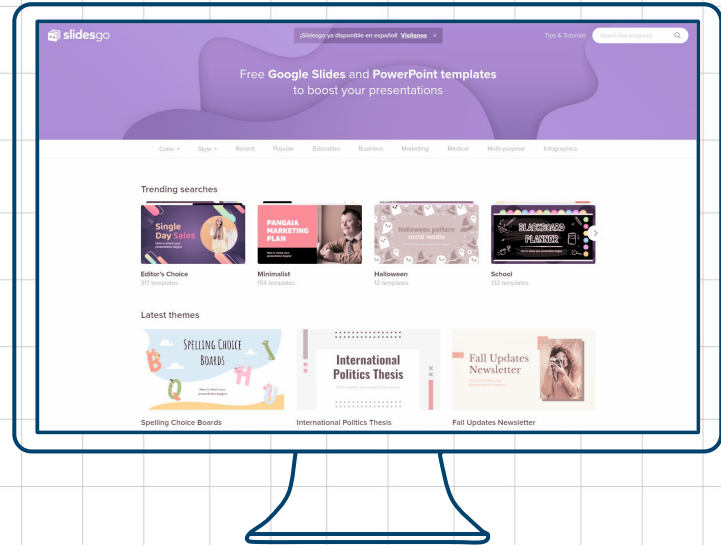
The closest planet to the Sun and the smallest one in the Solar System—it's only a bit larger than the Moon

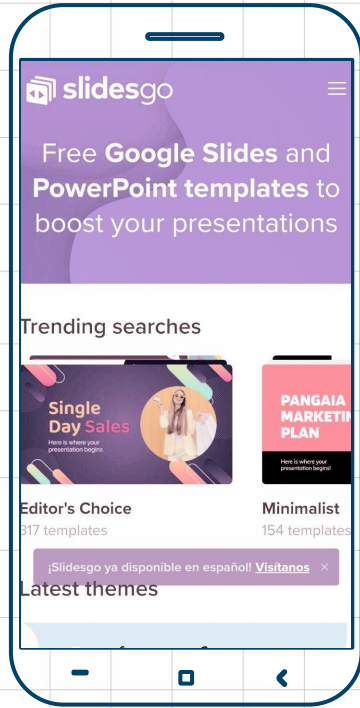
Bibliographical References

- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher

Desktop Software

You can replace the image on the screen with your own work. Just delete this one, add yours and center it properly



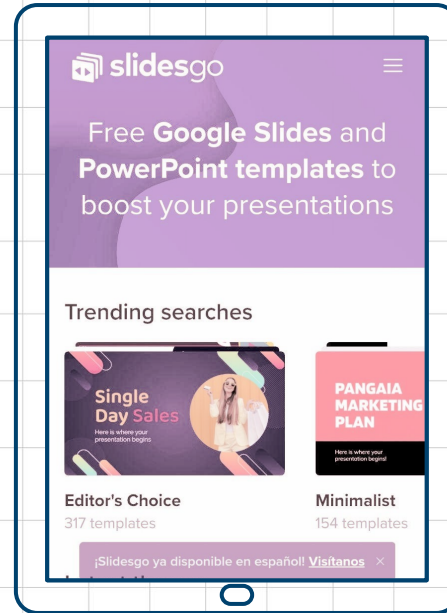


Mobile Web

You can replace the image on the screen with your own work. Just delete this one, add yours and center it properly

Tablet App

You can replace the image on the screen with your own work. Just delete this one, add yours and center it properly



Thanks!



Do you have any questions?
addyouremail@freepik.com
+91 620 421 838 yourcompany.com

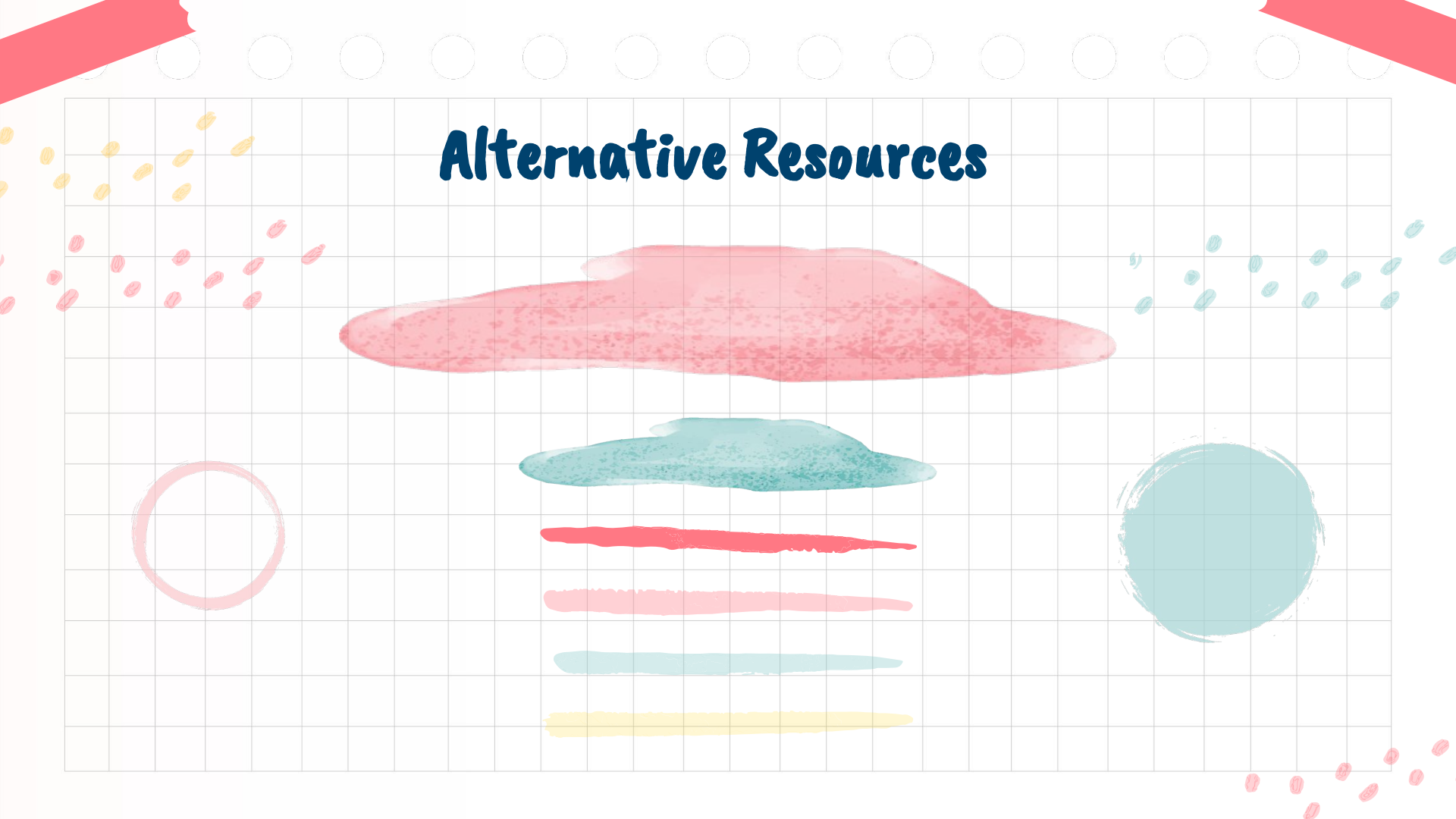
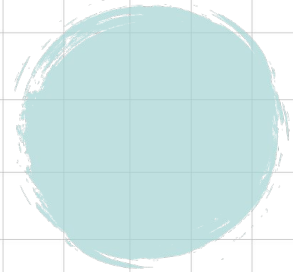
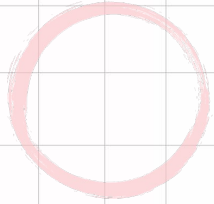
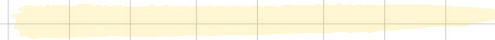
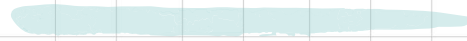
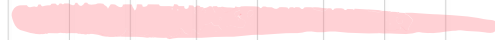
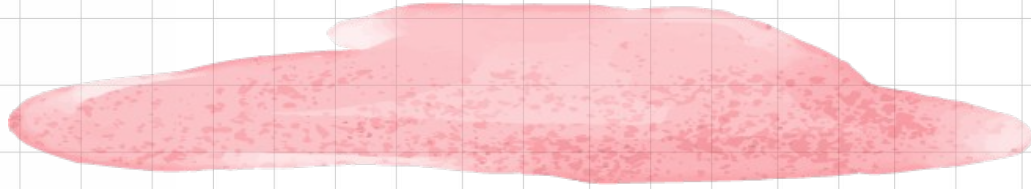
CREDITS: This presentation template was created by
Slidesgo, including icons by **Flaticon**, and infographics &
images by **Freepik**.

Please keep this slide for attribution.

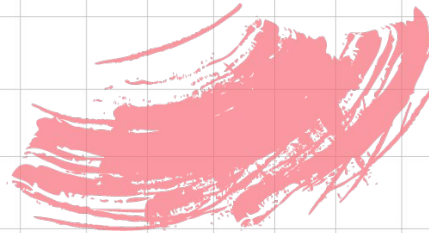
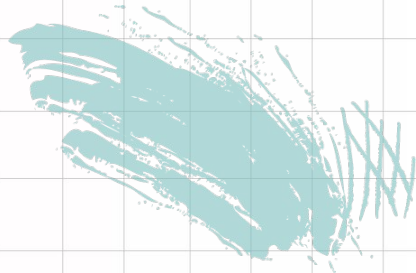
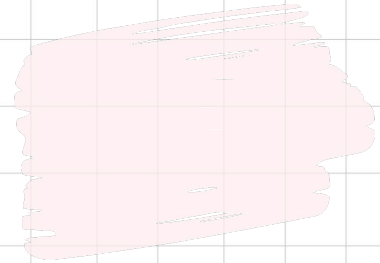
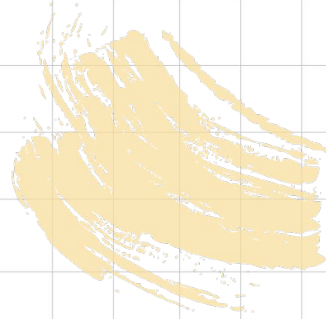
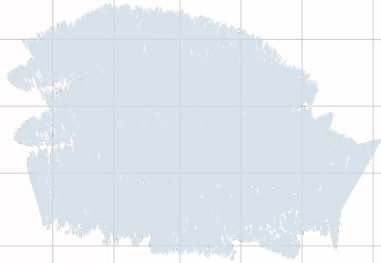
College Icon Pack



Alternative Resources



Alternative Resources



Alternative Resources

Photo

- Teacher and kids hugging outdoors close-up
- Kids hugging their teacher
- Teacher smelling a bouquet of flowers with copy space
- Teacher standing next to a blackboard with copy space
- Happy teacher helping her students




Resources



Photos

- Beautiful young female with paper cup in office
- Teacher writing on white board
- Portrait female teacher in class
- Daughter and mom hugging on couch
- Black father feeding son with croissant
- Smiley teacher standing in classroom
- Teacher hugging a student with copy space

Vectors

- Business infographic with different sheets of paper
 - Instagram stories template pattern
 - Abstract brush stroke pastel pattern
 - Abstract brush stroke pattern
 - Schedule with nice notes
- 

Instructions for use (free users)

In order to use this template, you must credit [Slidesgo](#) by keeping the Thanks slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Delete the “Thanks” or “Credits” slide.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

<https://slidesgo.com/faqs> and <https://slidesgo.com/slidesgo-school>

Instructions for use (premium users)

In order to use this template, you must be a Premium user on [Slidesgo](#).

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.
- Hide or delete the “Thanks” slide and the mention to Slidesgo in the credits.
- Share this template in an editable format with people who are not part of your team.

You are not allowed to:

- Sublicense, sell or rent this Slidesgo Template (or a modified version of this Slidesgo Template).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

<https://slidesgo.com/faqs> and <https://slidesgo.com/slidesgo-school>

Fonts & colors used

This presentation has been made using the following fonts:

Caveat Brush

(<https://fonts.google.com/specimen/Caveat+Brush>)

Raleway

(<https://fonts.google.com/specimen/Raleway>)

#595959

#ff7884

#ffb5bd

#ffc660

#00426e

#81c7c7

#003d6a

#a8bfce

Stories by Freepik

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out [How it Works](#).



Pana



Amico



Bro



Rafiki

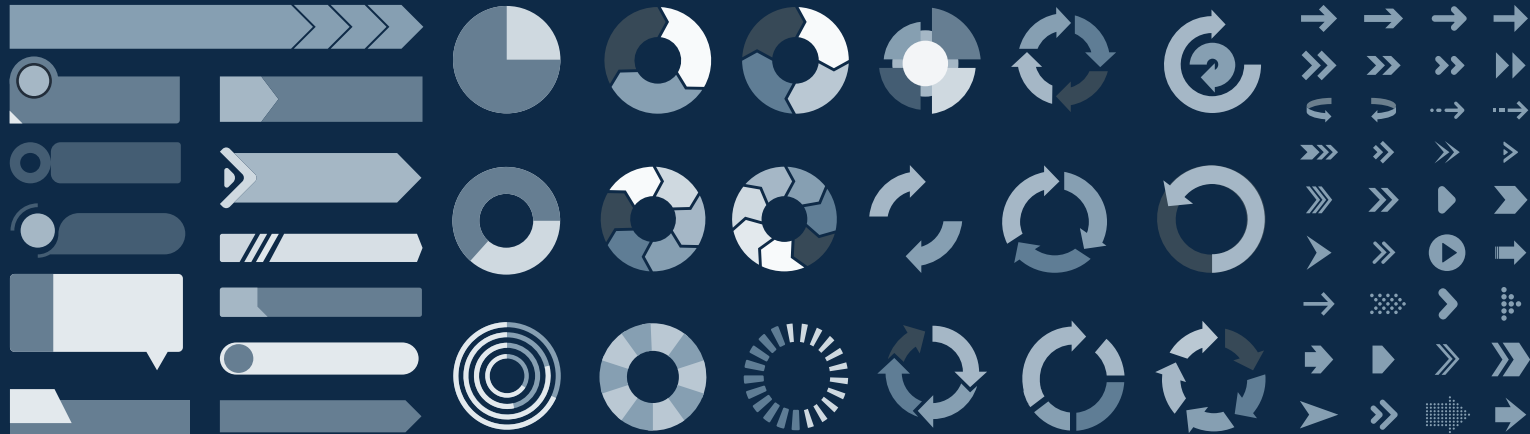


Cuate

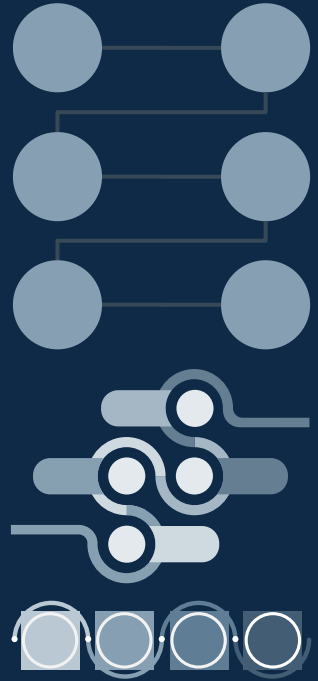
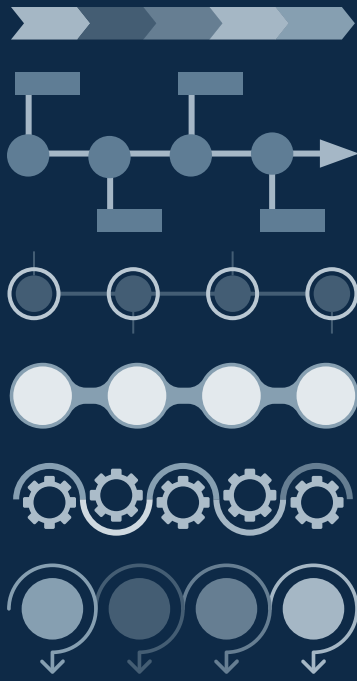
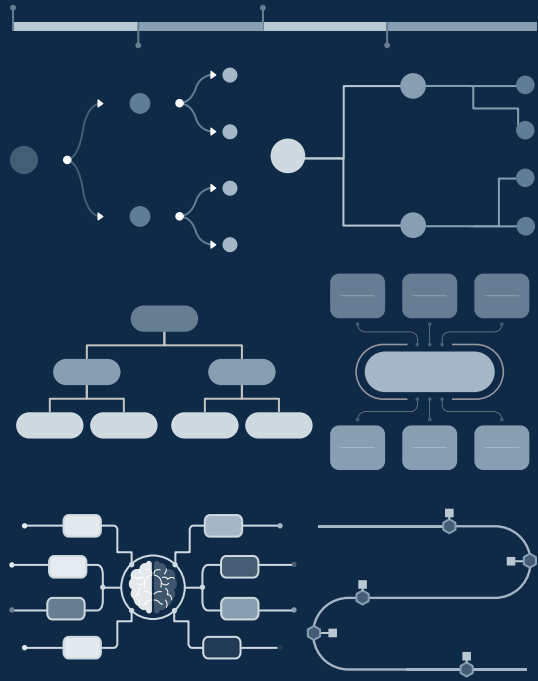
Use our editable graphic resources...

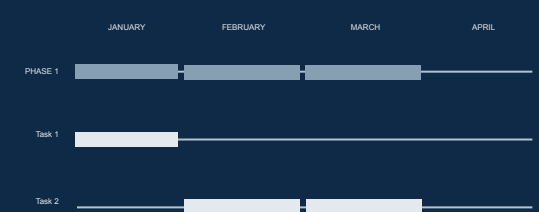
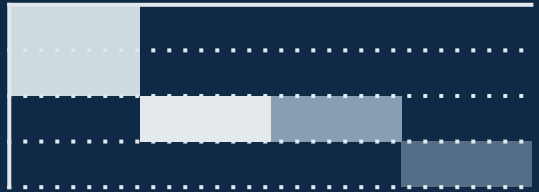
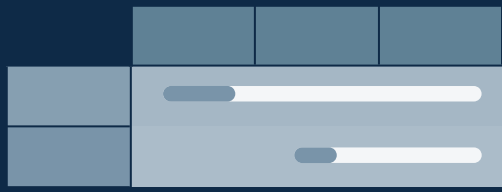
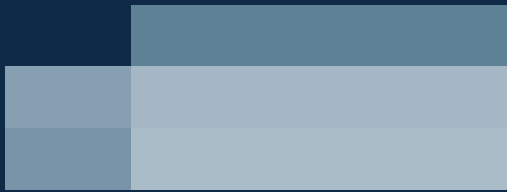
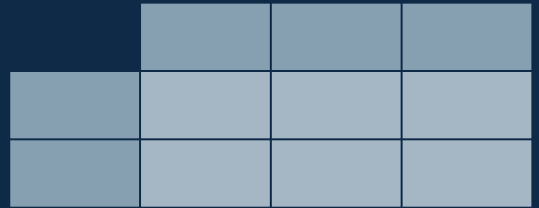
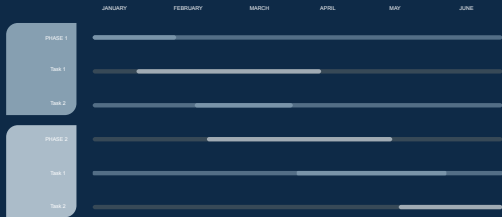
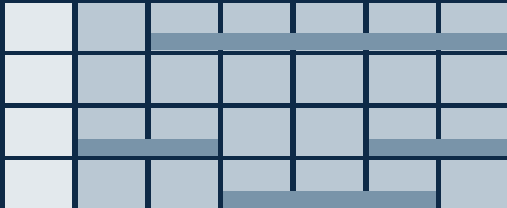
You can easily [resize](#) these resources without losing quality. To [change the color](#), just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want.

Group the resource again when you're done. You can also look for more [infographics](#) on [Slidesgo](#).

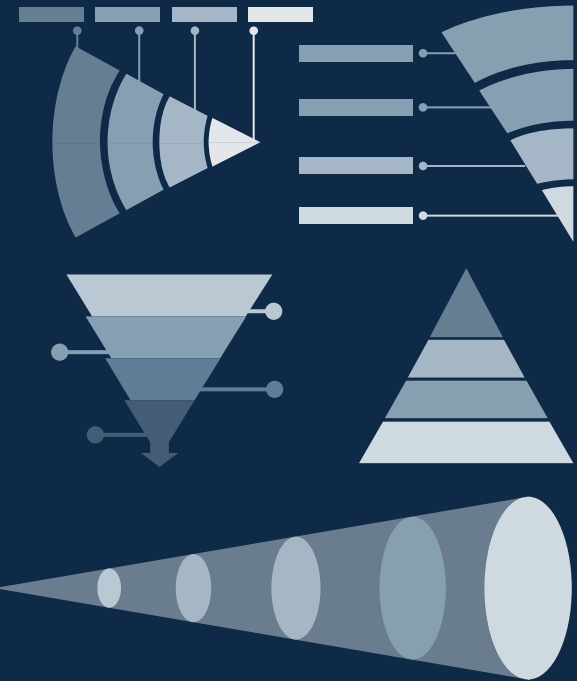
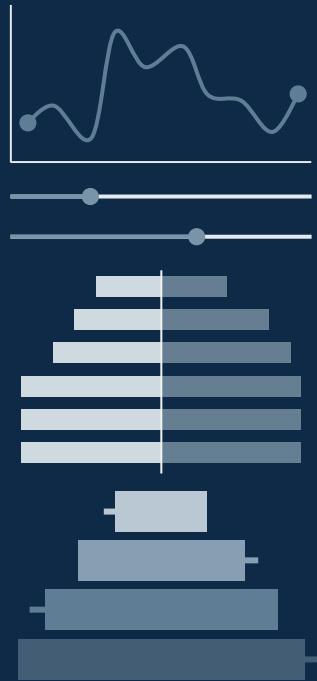
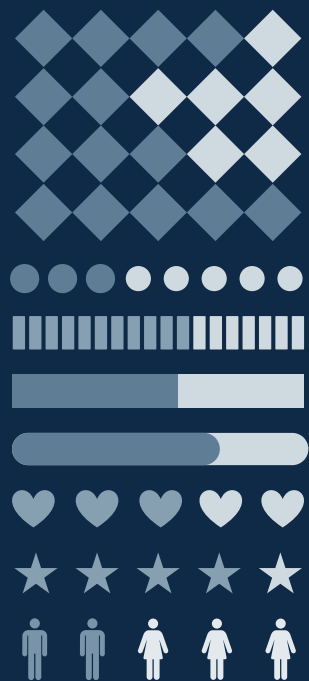
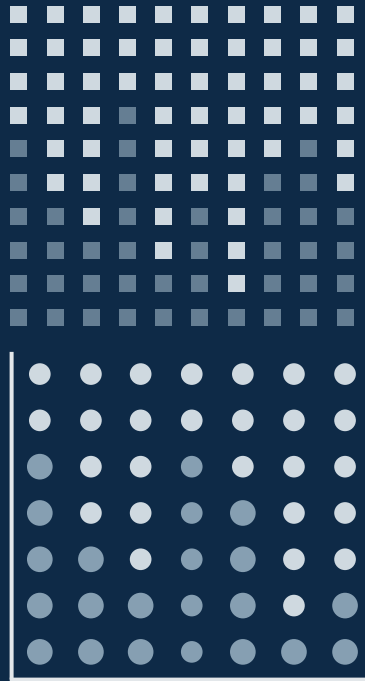












...and our sets of editable icons

You can resize these icons without losing quality.

You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.

In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.



Educational Icons



Medical Icons



Business Icons



Teamwork Icons



Creative Process Icons



Performing Arts Icons



Nature Icons



SEO & Marketing Icons



