

SC1015

Mini-Project

(Anime Score Prediction)

A133 Team 5

- Pang Xiu Quan, Lee Jun Yang, Ang Yu Quan

Motivation

01

Over 10,000 anime shows available over more than 20 years

02

Make the process of choosing which anime to watch more efficient and accurate



Problem statement:

How can we help viewers discover new and enjoyable anime series?








Dataset used

Top Anime

Top > Anime > Top Anime

All Anime Top Airing Top Upcoming Top TV Series Top Movies Top OVAs Top ONAs Top Specials Most Popular Most Favored

Top Anime Series Updated twice a day. (How do we rank shows?) [Next 50 >](#)

Rank	Title	Score	Your Score	Status
1	 "Oshi no Ko" PV TV (11 eps) Apr 2023 - 327,628 members	★ 9.19	★ N/A	Add to list
2	 Fullmetal Alchemist: Brotherhood TV TV (64 eps) Apr 2009 - Jul 2010 3,138,821 members	★ 9.10	★ N/A	Add to list
3	 Shingeki no Kyojin: The Final Season - Kanketsu-hen PV Special (2 eps) Mar 2023 - 2023 397,736 members	★ 9.09	★ N/A	Add to list
4	 Steins;Gate TV TV (24 eps) Apr 2011 - Sep 2011 2,413,139 members	★ 9.08	★ N/A	Add to list
5	 Bleach: Sennen Kessen-hen PV TV (13 eps) Oct 2022 - Dec 2022 421,795 members	★ 9.08	★ N/A	Add to list
6	 Gintama° TV TV (51 eps) Apr 2015 - Mar 2016 587,485 members	★ 9.07	★ N/A	Add to list
7	 Kaguya-sama wa Kokurasetai: Ultra Romantic TV TV (13 eps) Apr 2022 - Jun 2022 793,749 members	★ 9.06	★ N/A	Add to list

MyAnimeList dataset:

- Titles
- Synopses
- Genres
- Mean score
- Etc.

Retrieved from MyAnimeList.net API

Data Preparation & Cleaning

```
anime_details = []
headers = {'Authorization': 'Bearer eyJ0eXAiOiJKV1QiLCJhbGciOiJSUzI1NiIs
params = {
    'fields': 'id,title,main_picture,alternative_titles,start_date,end_d
    'limit' : 1
}

for anime_id in anime_ids:
    try:
        url = f'https://api.myanimelist.net/v2/anime/{anime_id}'
        response = requests.get(url, headers=headers, params=params)
        data = json.loads(response.text)
        anime_details.append(data)

        # Wait 1 seconds for the next request interval
        time.sleep(1)
    except Exception as e:
        print(f"Error retrieving data for anime ID {anime_id}: {e}")

df = pd.json_normalize(anime_details)
df.to_csv('anime_details.csv', index=False)
```

01 API GET dataset

02 Discard unnecessary columns
and Animes without Mean

03 Filled in missing values

04 Opened 'Genres' columns to
create individual columns

Data Preparation & Cleaning

```
df.drop(['start_date', 'end_date', 'synopsis',  
        'created_at', 'updated_at',  
        'average_episode_duration', 'pictures',  
        'background', 'related_anime',  
        'related_manga', 'recommendations',  
        'main_picture.medium', 'main_picture.large',  
        'alternative_titles.synonyms',  
        'alternative_titles.en', 'alternative_titles.ja',  
        'broadcast.day_of_the_week', 'broadcast.start_time'],  
        axis=1, inplace=True)  
df = df.dropna(subset=['mean', 'rank',  
                      'start_season.year',  
                      'start_season.season'])  
df = df[df['start_season.year'] >= 1999]
```

02

Discard unnecessary columns
and rows

03

Filled in missing values

04

Opened 'Genres' columns to
create individual columns

Data Preparation & Cleaning

```
df["genres"].fillna(value = "[{'id': -1, 'name': 'no genre'}]", inplace = True)
df["source"].fillna(value = "unknown", inplace = True)
df["rating"].fillna(value = "no rating", inplace = True)
df['studios'] = df['studios'].replace("[]", "[{'id': -1, 'name': 'no studio'}]")
```

03

Filled in missing values

04

Opened 'Genres' columns to
create individual columns

Data Preparation & Cleaning

```
# Create an empty list to store all genre names
genre_names = []

df['genres'] = df['genres'].str.replace("'", '')

# Loop through each row in the 'genres' column
for genres in df['genres']:
    # Check if the value in the 'genres' column is NaN or float
    if pd.isna(genres):
        # Convert NaN to an empty list
        genres_list = []
    elif isinstance(genres, float):
        # Convert float to an empty list
        genres_list = []
    else:
        # Parse the string into a list of dictionaries
        # using the JSON module
        genres_list = json.loads(genres)
    # Loop through each dictionary in the list and
    # append the 'name' key to the 'genre_names' list
    for genre in genres_list:
        genre_names.append(genre['name'])

# Get a set of unique genre names
unique_genres = set(genre_names)

# Loop through each unique genre name
for genre in unique_genres:
    # Create a new column with the genre name and populate
    # it with True or False values depending on whether the
    # genre is present in the 'genres' column for each row
    df[genre] = df['genres'].apply(lambda x: genre in
                                   [g['name'] for g in json.loads(x)]
                                   if isinstance(x, str) else False)

# Drop the original 'genres' column
df.drop('genres', axis=1, inplace=True)
```

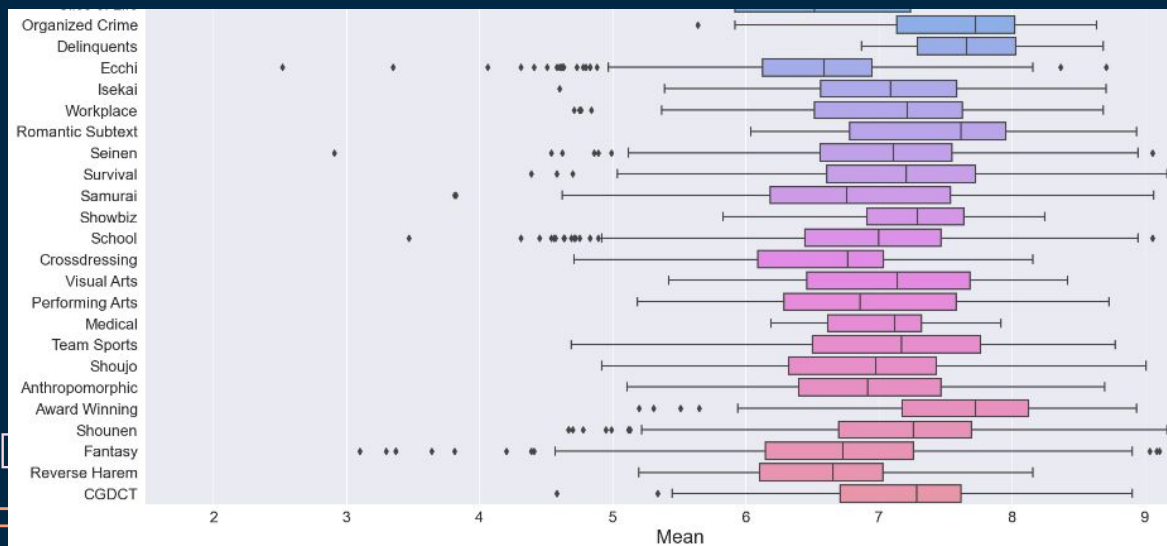
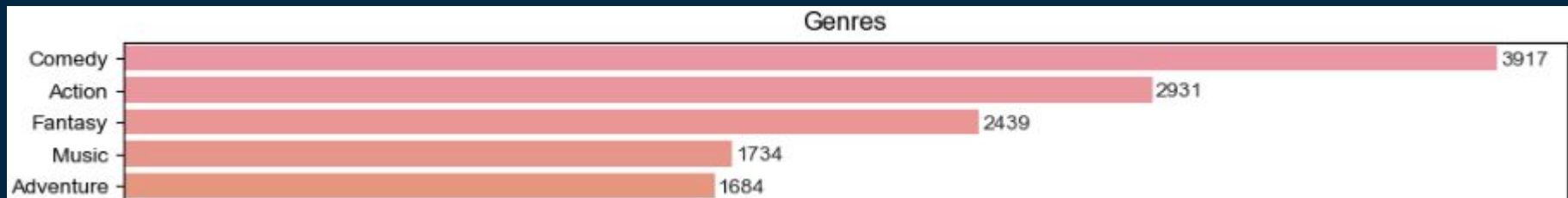
04

Opened 'Genres' columns to
create individual columns

EDA

Mean	Average score rated by the viewers	Rating	Content rating indicting appropriate age range for the anime
NSFW	Whether the anime is "Not Safe For Work" or not	Studio	The animation studios that produced the anime
Media	The type of media that the anime belongs to	Season	The year which the anime started airing
Genres	The different categories or classification that are assigned to the anime	Year	The season which the anime started airing
Source	The medium which the anime is adapted from	Watch rate	Percentage of people who are watching or finished the anime
Popularity	Relative popularity of the anime compared to other anime titles on the website	id	Unique number of the anime

EDA (Genres)



Genre and mean correlation:

Adult Cast: 0.16012616436645202

Otaku Culture: 0.05943837427487703

Music: -0.20723176706039803

Educational: -0.002162207785528912

no genre: -0.03636243192231978

Time Travel: 0.06716210500828824

Mystery: 0.1461927349626088

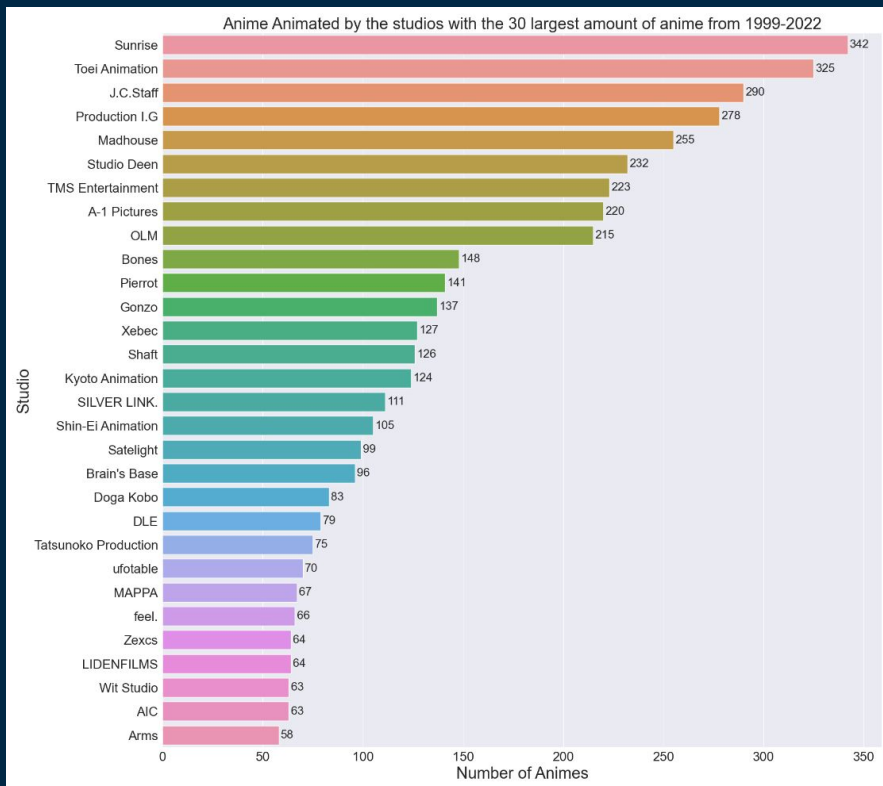
Iyashikei: 0.13965859170142125

Super Power: 0.0693173813036846

Combat Sports: 0.024271510066078278

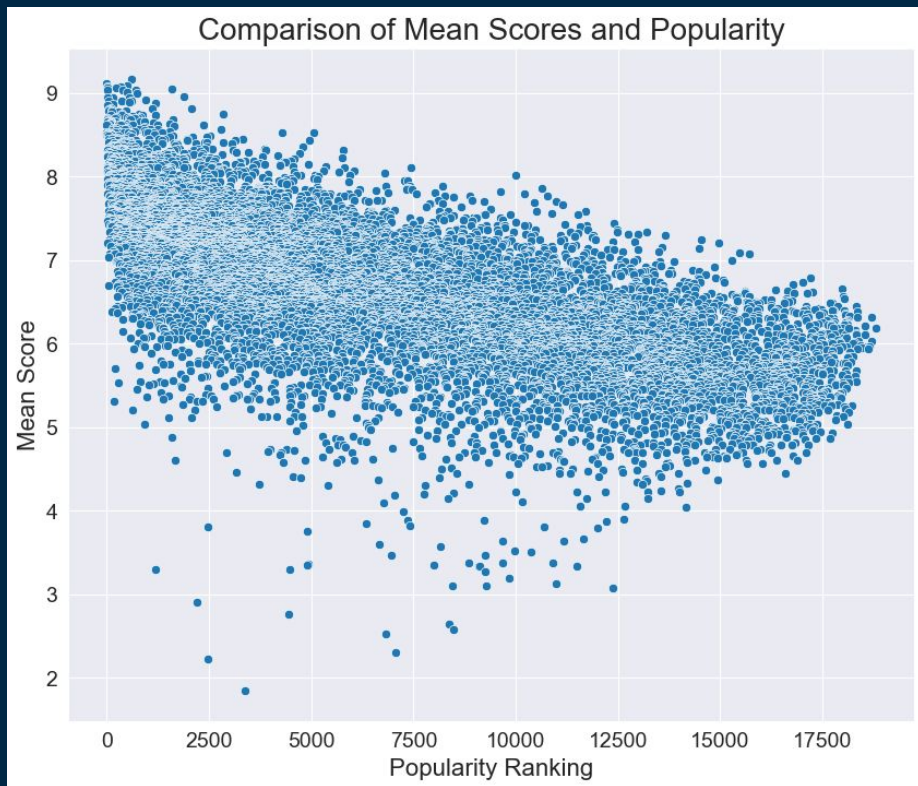
Mahou Shoujo: 0.051219067542870686

EDA (Studios)

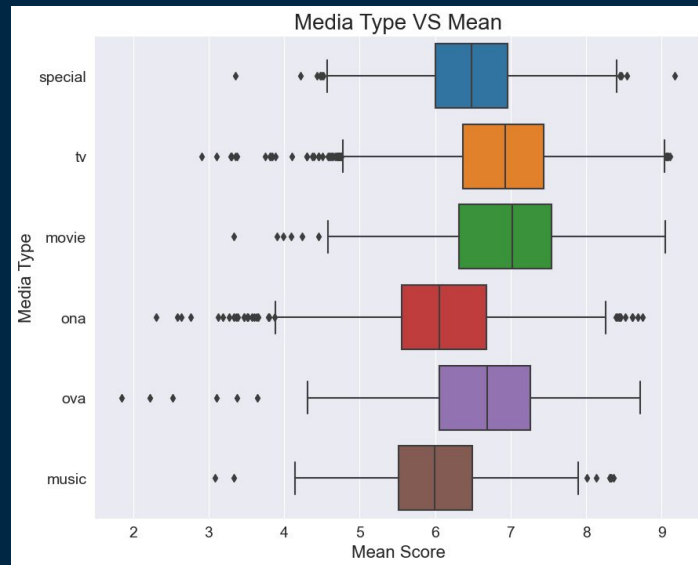
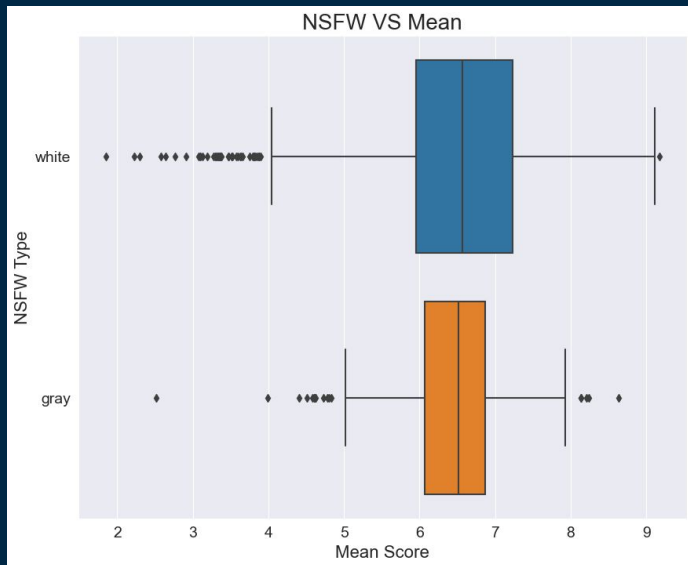


Madhouse: 7 %
 MAPPA: 6 %
 Sunrise: 6 %
 Bandai Namco Pictures: 5 %
 A-1 Pictures: 5 %
 Wit Studio: 5 %
 Kyoto Animation: 5 %
 Shaft: 5 %
 Artland: 5 %
 Bones: 4 %
 Pierrot: 4 %
 ufotable: 4 %
 Production I.G: 3 %
 Studio Deen: 3 %
 Kinema Citrus: 3 %
 Brain's Base: 3 %
 TMS Entertainment: 2 %
 CloverWorks: 2 %
 CoMix Wave Films: 2 %
 Studio Ghibli: 2 %
 David Production: 2 %
 Gainax: 2 %
 Shuka: 2 %
 White Fox: 1 %
 Studio LAN: 1 %
 Studio Bind: 1 %
 OLM: 1 %
 Toei Animation: 1 %
 Tatsunoko Production: 1 %
 Khara: 1 %
 B.CMAY PICTURES: 1 %
 Trigger: 1 %
 Studio Chizu: 1 %
 Studio VOLN: 1 %
 Shin-Ei Animation: 1 %
 J.C.Staff: 1 %

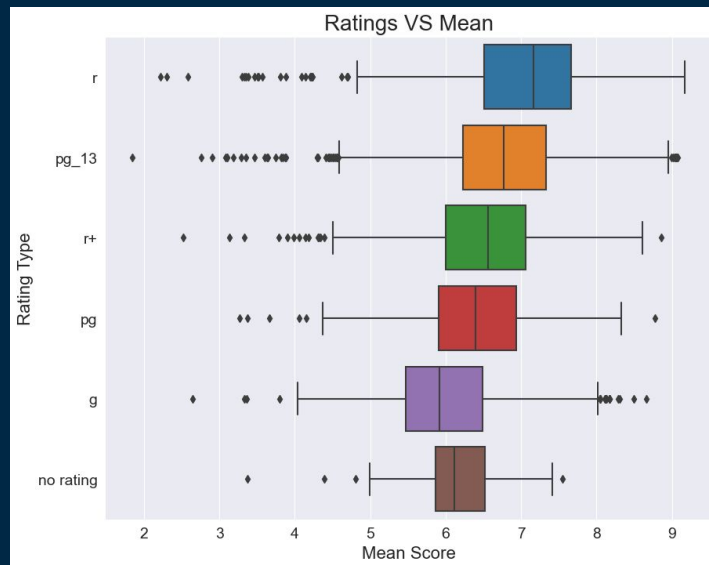
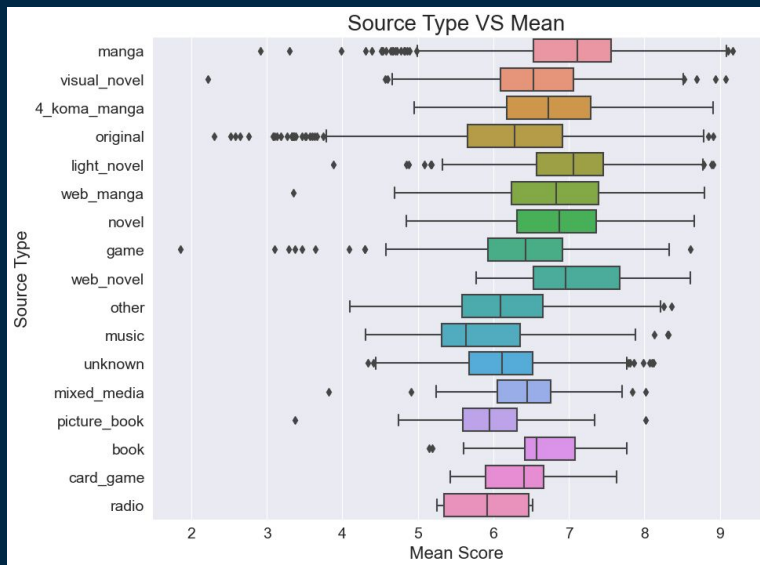
EDA (Popularity)



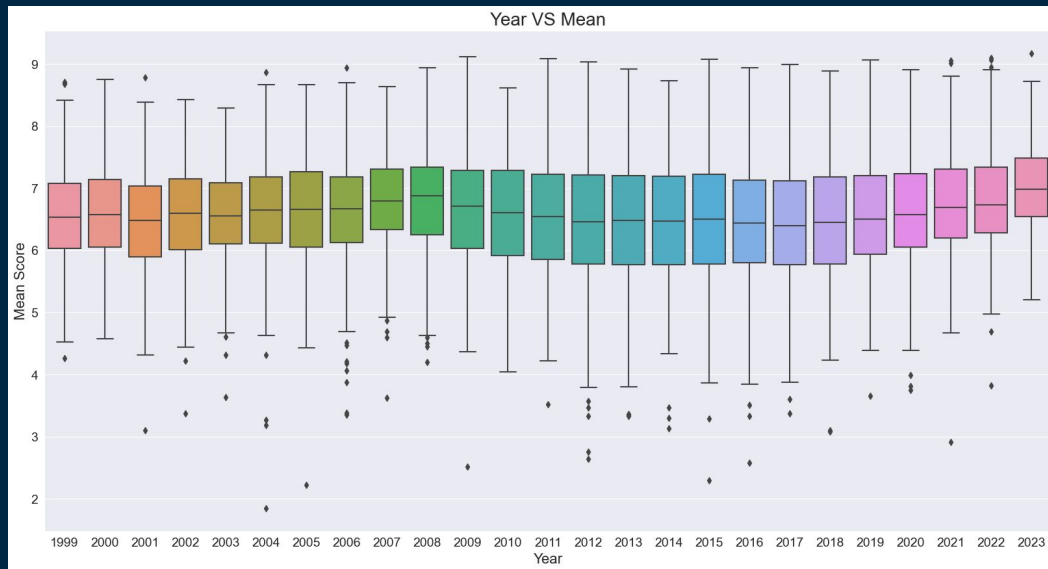
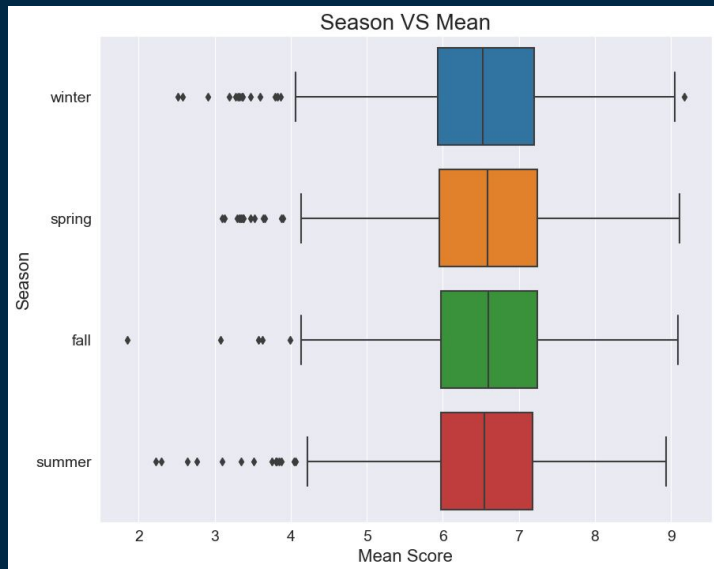
EDA (others)



EDA (others)



EDA (others)



Tools and technique used from this course



EDA

Machine Learning

Others

- Different plots

- Linear Regression
- Classification Tree

- -

New Tools and technique used



EDA

- Learned how to use lambda function
- Use groupby function



Machine Learning

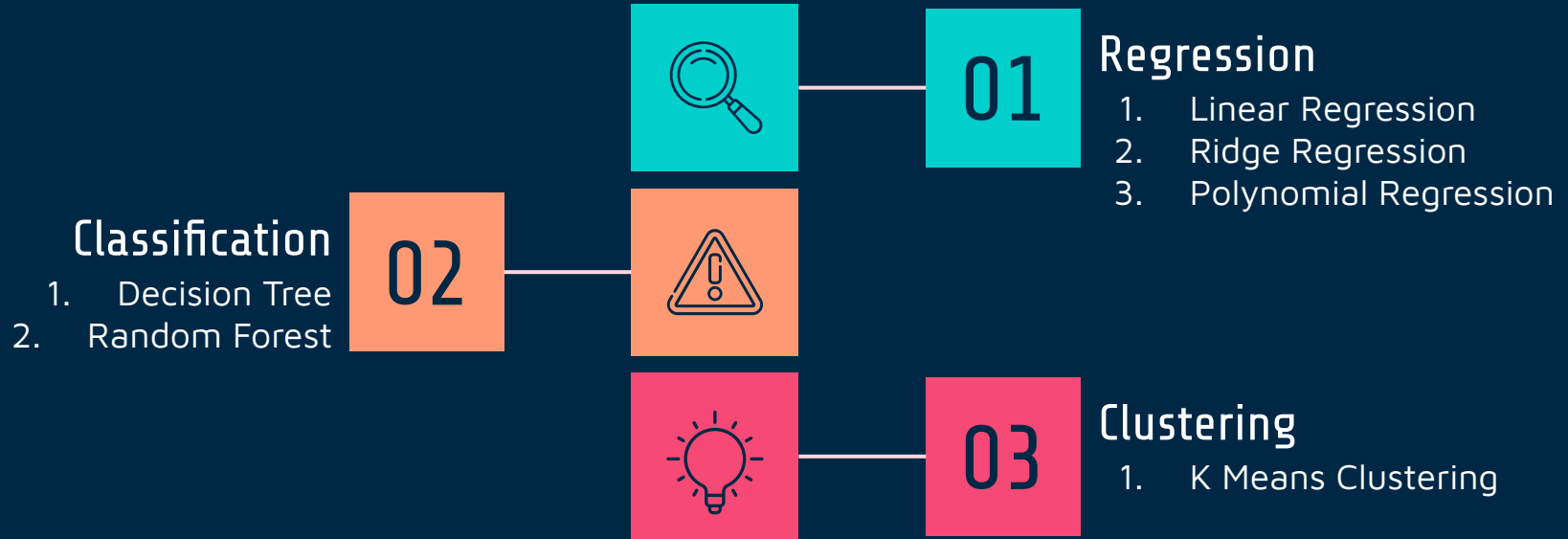
- Ridge and Polynomial Regression
- Random Forest
- K means clustering



Others

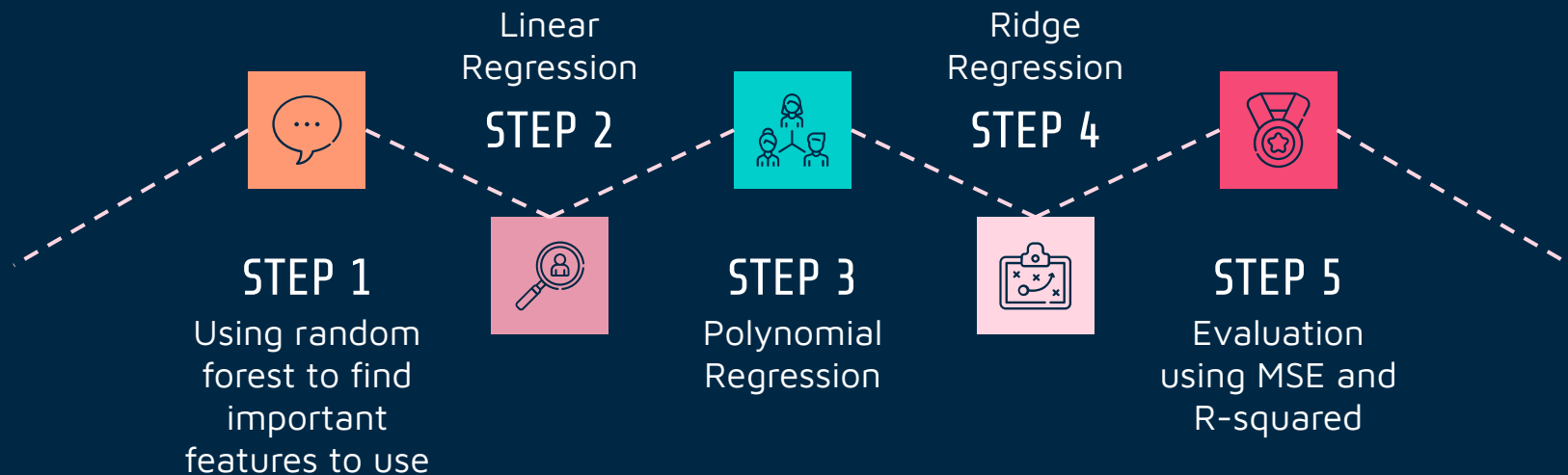
- Make use of one hot encoding to change categorical into numerical data
- Accessing API
- Using GraphViz

Machine Learning Models



Regression

Step-by-step Summary



Rank of features from Random Forest

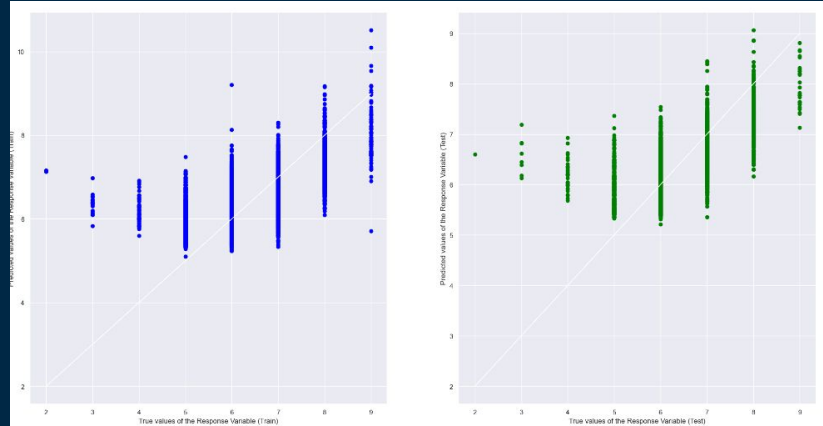
- Populairty
- Statistics number list user
- Number of list user
- Number of Scoring users
- Watch rate
- Statistics plan to watch
- Start season year
- Rating PG 13
- Start season Fall
- Number of episodes

Using GINI importance

- Checks which feature will reduce the impurity more

Choosing which regression model

Linear Regression



Goodness of Fit of Model
Explained Variance (R^2)
Mean Squared Error (MSE)

Test Dataset
: 0.442470642255121
: 0.4777798720589073

Choosing which regression model

Ridge Regression

Goodness of Fit of Model

Explained Variance (R^2)

Mean Squared Error (MSE)

Test Dataset

: 0.4590407858180773

: 0.4551671061261718

Choosing which regression model

Polynomial Regression

Goodness of Fit of Model

Explained Variance (R^2)

Mean Squared Error (MSE)

Test Dataset

: 0.5145862214772292

: 0.4093972896990084

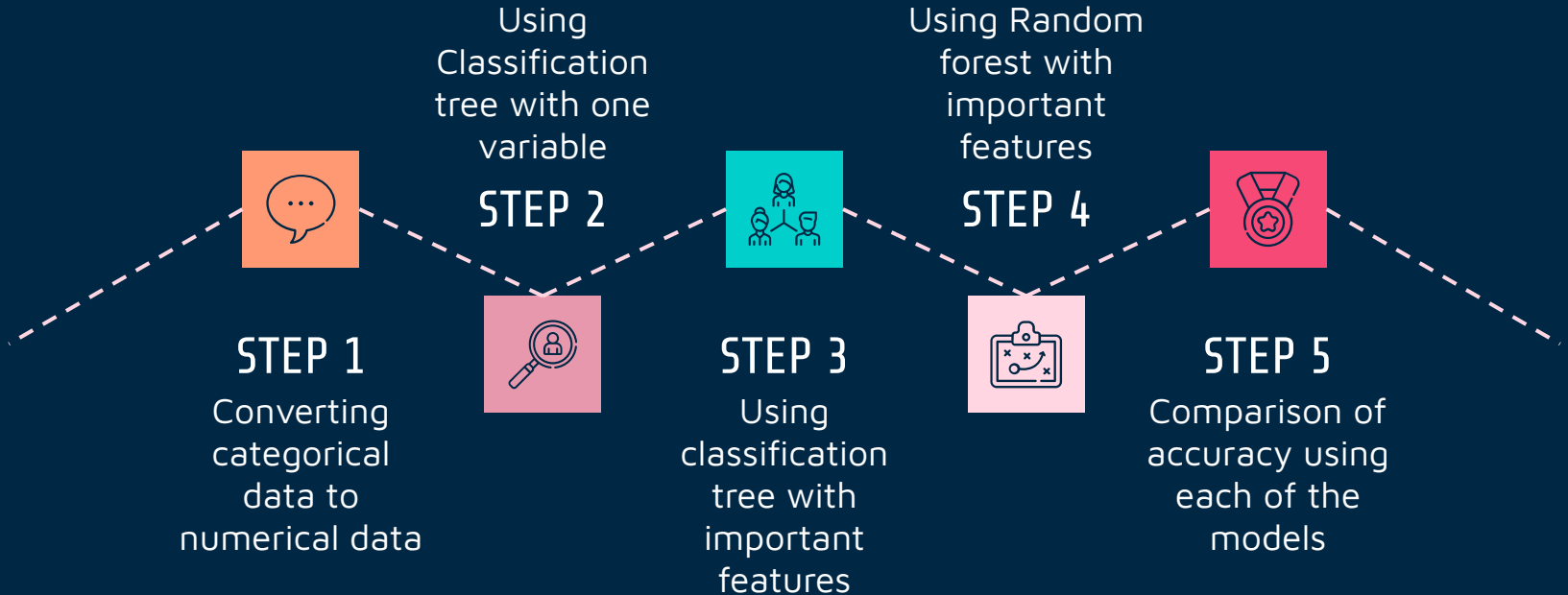
Choosing which regression model

Comparing MSE and R^2

Best Regression Model is Polynomial Regression

Machine Learning – Classification

Step-by-step Summary



Converting from categorical to numeric

Example of conversion

Features that are being converted

1. Rating of anime
2. Source of anime story
3. Studio which created the anime
4. Season in which the anime is produced

```
# Conversion of studio_only to num
df1['studio_only'] = df1['studio_only'].replace(['no studio'],0)
df1['studio_only'] = df1['studio_only'].replace(['Sunrise'],1)
df1['studio_only'] = df1['studio_only'].replace(['Toei Animation'],2)
df1['studio_only'] = df1['studio_only'].replace(['J.C.Staff'],3)
df1['studio_only'] = df1['studio_only'].replace(['Production I.G'],4)
df1['studio_only'] = df1['studio_only'].replace(['Madhouse'],5)
df1['studio_only'] = df1['studio_only'].replace(['Studio Deen'],6)
df1['studio_only'] = df1['studio_only'].replace(['A-1 Pictures'],7)
df1['studio_only'] = df1['studio_only'].replace(['OLM'],8)
df1['studio_only'] = df1['studio_only'].replace(['TMS Entertainment'],9)
df1['studio_only'] = df1['studio_only'].replace(['Bones'],10)
df1['studio_only'] = df1['studio_only'].replace(['Pierrot'],11)
df1['studio_only'] = df1['studio_only'].replace(['Gonzo'],12)
df1['studio_only'] = df1['studio_only'].replace(['Shaft'],13)
df1['studio_only'] = df1['studio_only'].replace(['Kyoto Animation'],14)
df1['studio_only'] = df1['studio_only'].replace(['Xebec'],15)
df1['studio_only'] = df1['studio_only'].replace(['SILVER LINK'],16)
df1['studio_only'] = df1['studio_only'].replace(['Brain's Base'],17)
df1['studio_only'] = df1['studio_only'].replace(['Satelight'],18)
df1['studio_only'] = df1['studio_only'].replace(['Doga Kobo'],19)
df1['studio_only'] = df1['studio_only'].replace(['Shin-Ei Animation'],20)
df1['studio_only'] = df1['studio_only'].replace(['feel'],21)
df1['studio_only'] = df1['studio_only'].replace(['ufotable'],22)
df1['studio_only'] = df1['studio_only'].replace(['Zexcs'],23)
df1['studio_only'] = df1['studio_only'].replace(['MAPPA'],24)
df1['studio_only'] = df1['studio_only'].replace(['LIDENFILMS'],25)
df1['studio_only'] = df1['studio_only'].replace(['Tatsunoko Production'],26)
df1['studio_only'] = df1['studio_only'].replace(['Wit Studio'],27)
df1['studio_only'] = df1['studio_only'].replace(['DLE'],28)
df1['studio_only'] = df1['studio_only'].replace(['AIC'],29)
df1['studio_only'] = df1['studio_only'].replace(['Gainax'],30)

df1['studio_only'] = pd.to_numeric(df1['studio_only'], errors='coerce').fillna(value=0).astype(int)
```

```
# Conversion of season to num
df1['start_season.season'] = df1['start_season.season'].replace(['spring'],1)
df1['start_season.season'] = df1['start_season.season'].replace(['winter'],2)
df1['start_season.season'] = df1['start_season.season'].replace(['fall'],3)
df1['start_season.season'] = df1['start_season.season'].replace(['summer'],4)
```

Converting from categorical to numeric

Using classification tree with one variable

```
Classification Accuracy(Train) : 0.41282722513089004  
Classification Accuracy(Test)  : 0.4031413612565445
```

Using classification tree with important features

```
Classification Accuracy on Train data: 1.0  
Classification Accuracy on Test data: 0.5041884816753927
```

Using random forest with important features

```
Train Model Accuracy: 1.0  
Prediction Accuracy: 0.6057591623036649
```

Conclusion

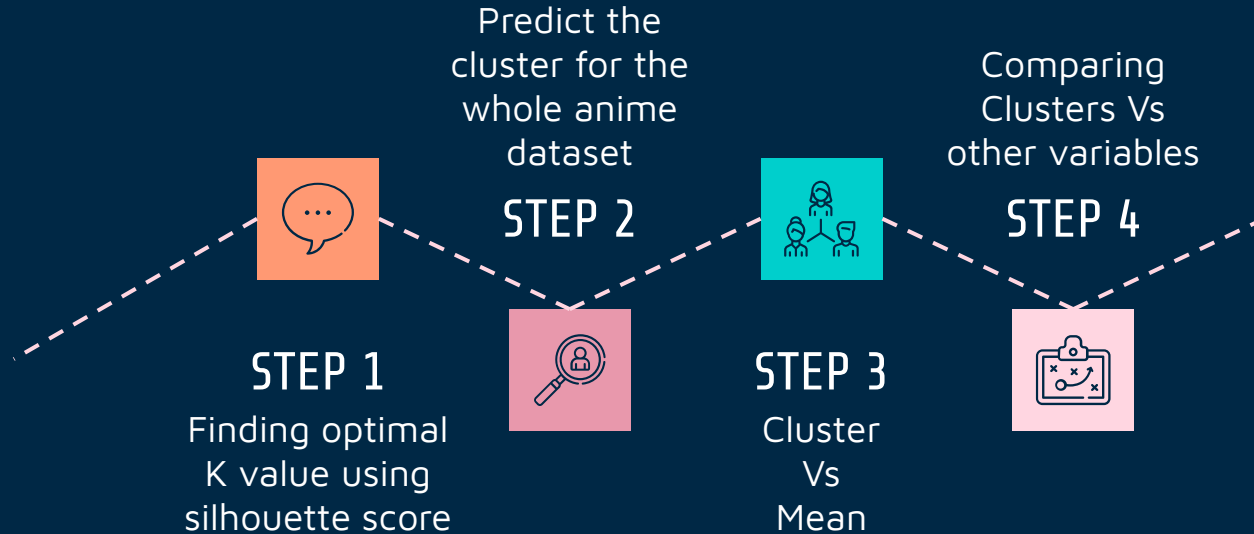
1. Random forest has the highest prediction accuracy (0.606)
2. Susceptible to bias and overfitting and can be limited by data availability
3. Given data is heavily skewed (Majority rating of 6 and 7)
4. Regression models may be more suitable than classification models due to the nature of data

Future fixes

1. Use resampling techniques such as oversampling or undersampling
2. Obtain more balanced dataset

Machine learning – Clustering

Step-by-step Summary



Finding the optimal K value using silhouette score

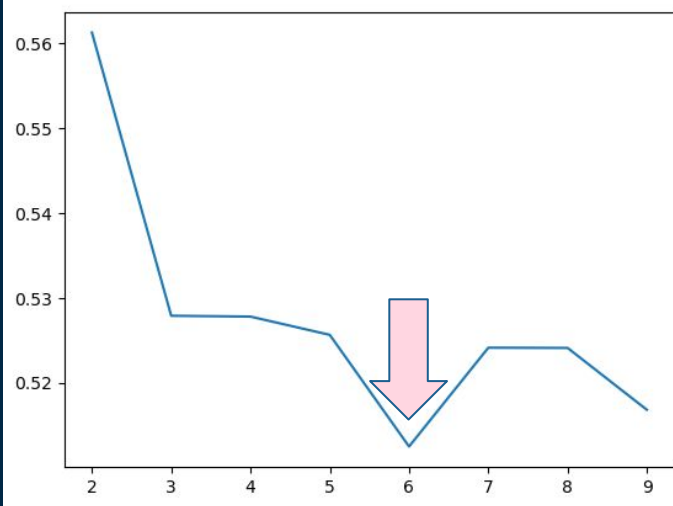
```
K = range(2, 10)
fits = []
score = []

for k in K:
    # train the model for current value of k on training data
    model = KMeans(n_clusters=int(k), random_state=0, n_init=10).fit(X_train_norm)

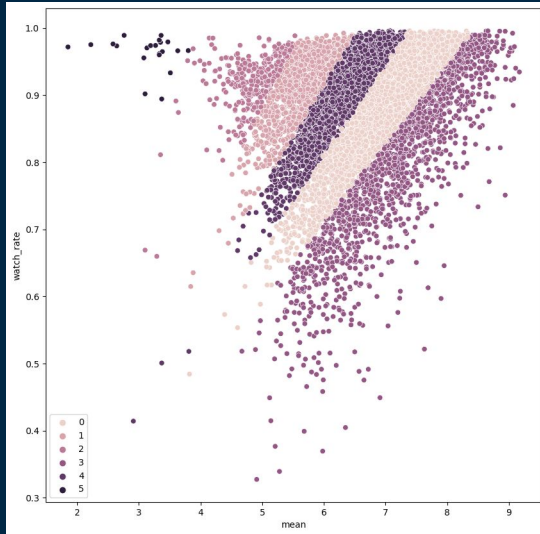
    # append the model to fits
    fits.append(model)

    # Append the silhouette score to scores
    score.append(silhouette_score(X_train_norm, model.labels_, metric='euclidean'))

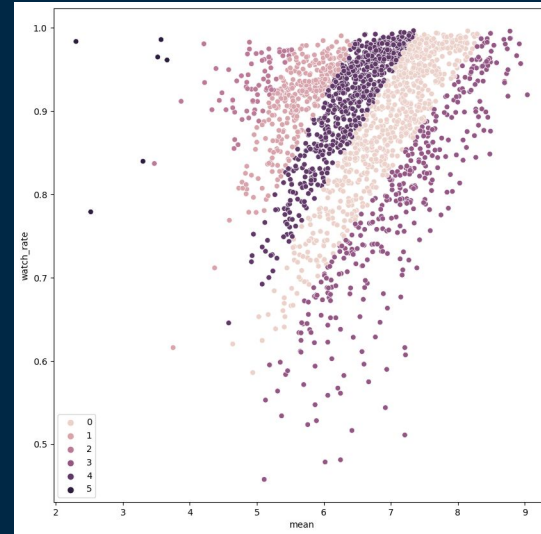
sb.lineplot(x = K, y = score)
```



Training dataset



Testing Dataset



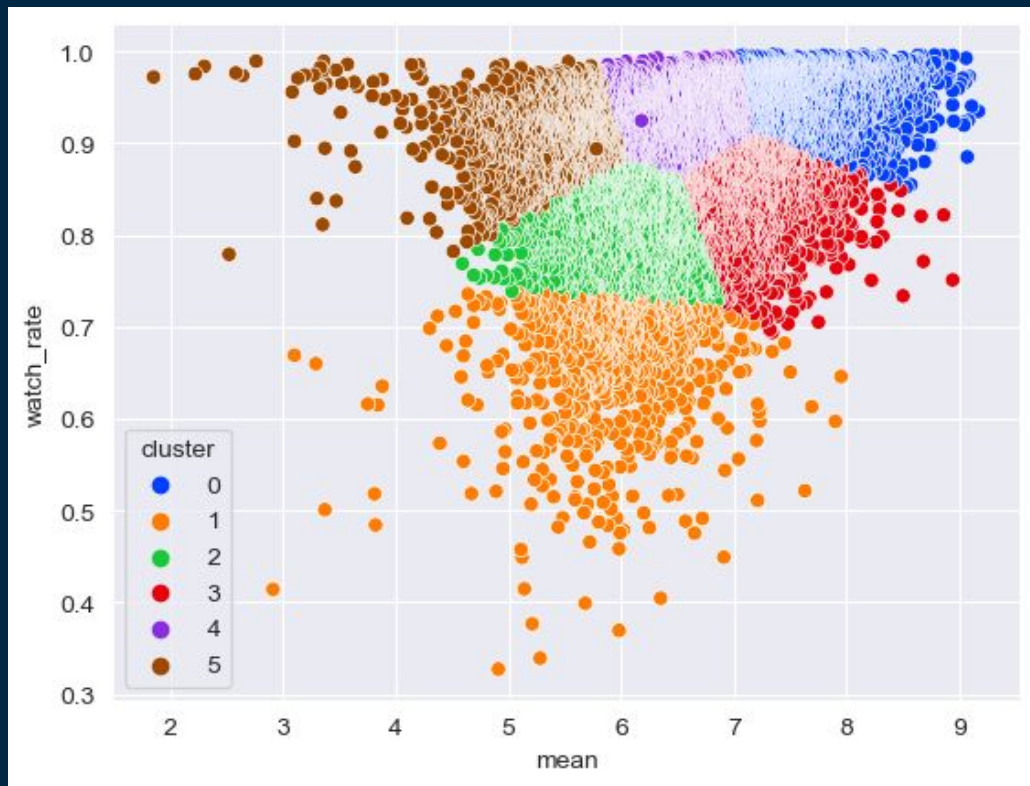
```
print('The Kmean clustering of training dataset with K = 6 has a silhouette score of',silhouette_score(X_train_norm, kmeans.labels_))
```

The Kmean clustering of training dataset with K = 6 has a silhouette score of 0.5124441148520357

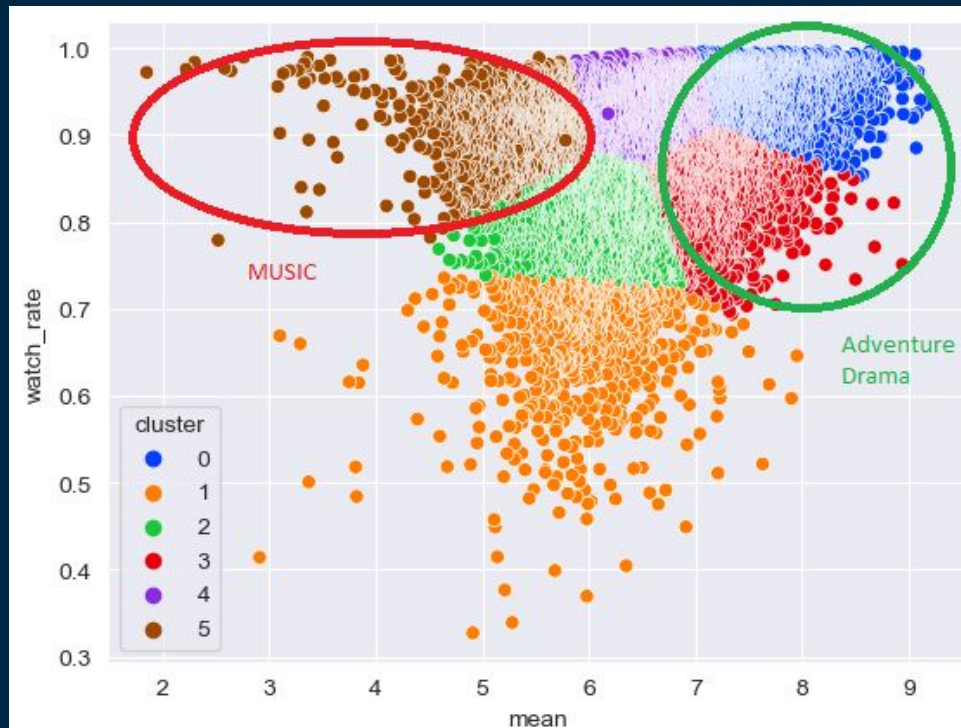
```
print('The Kmean clustering of test dataset with K = 6 has a silhouette score of',silhouette_score(X_test_norm, test_labels, metric='euclidean'))
```

The Kmean clustering of test dataset with K = 6 has a silhouette score of 0.512939922628873

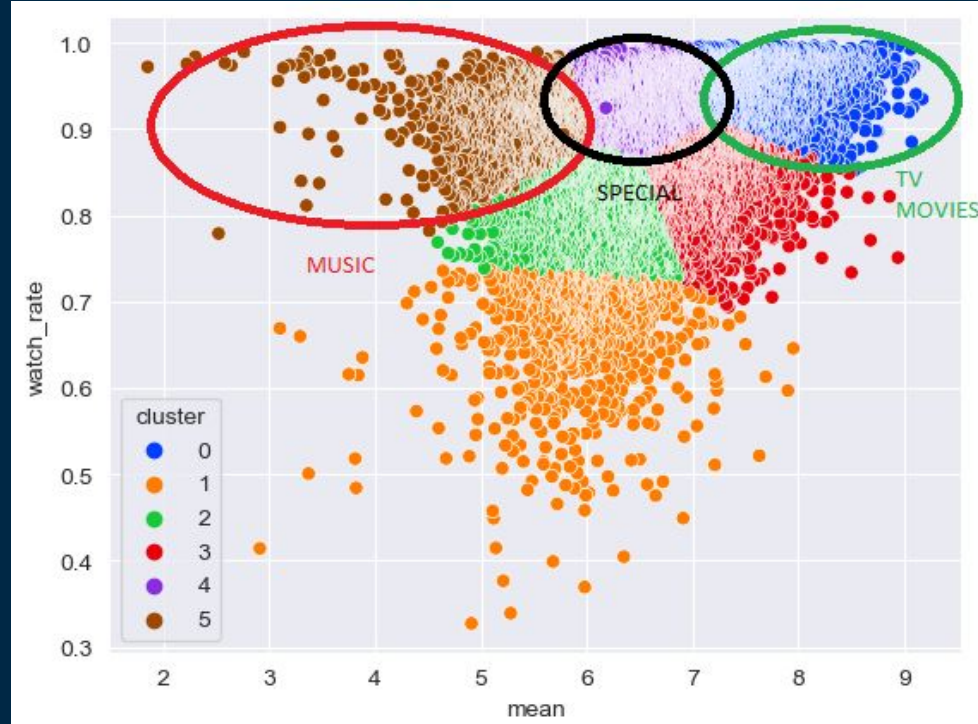
Clustered scatter plot of anime dataset



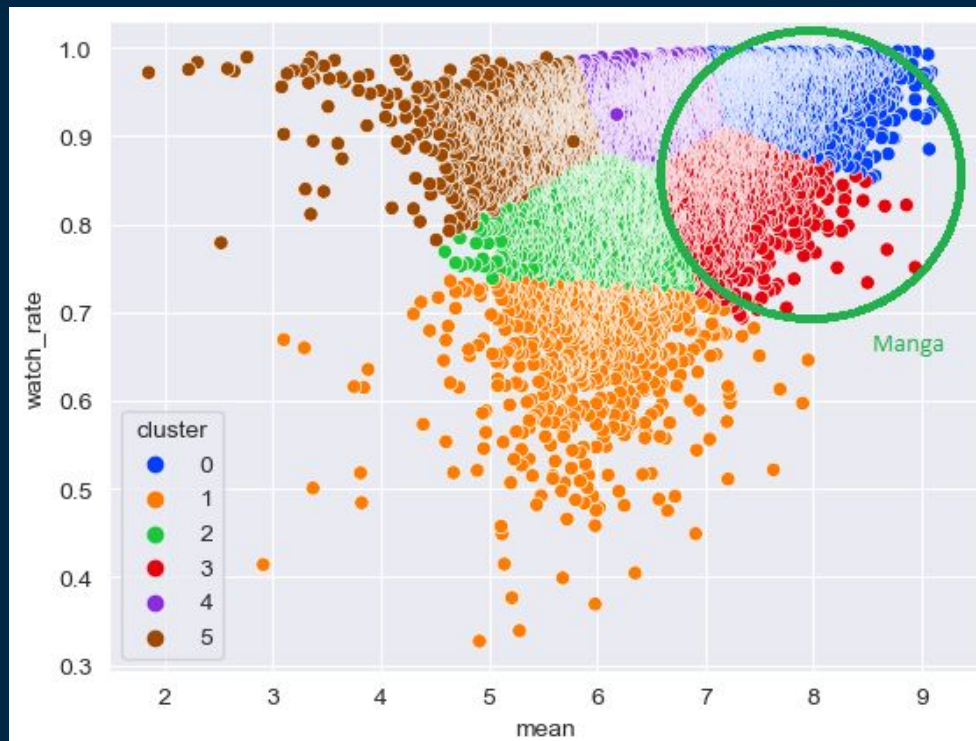
Clustered scatter plot of anime dataset (Genres)



Clustered scatterplot of anime dataset (Media type)



Clustered scatterplot of anime dataset (Media type)



Conclusion

01

Better model for regression: **Random forest**

Better model for Classification: **Polynomial regression**

02

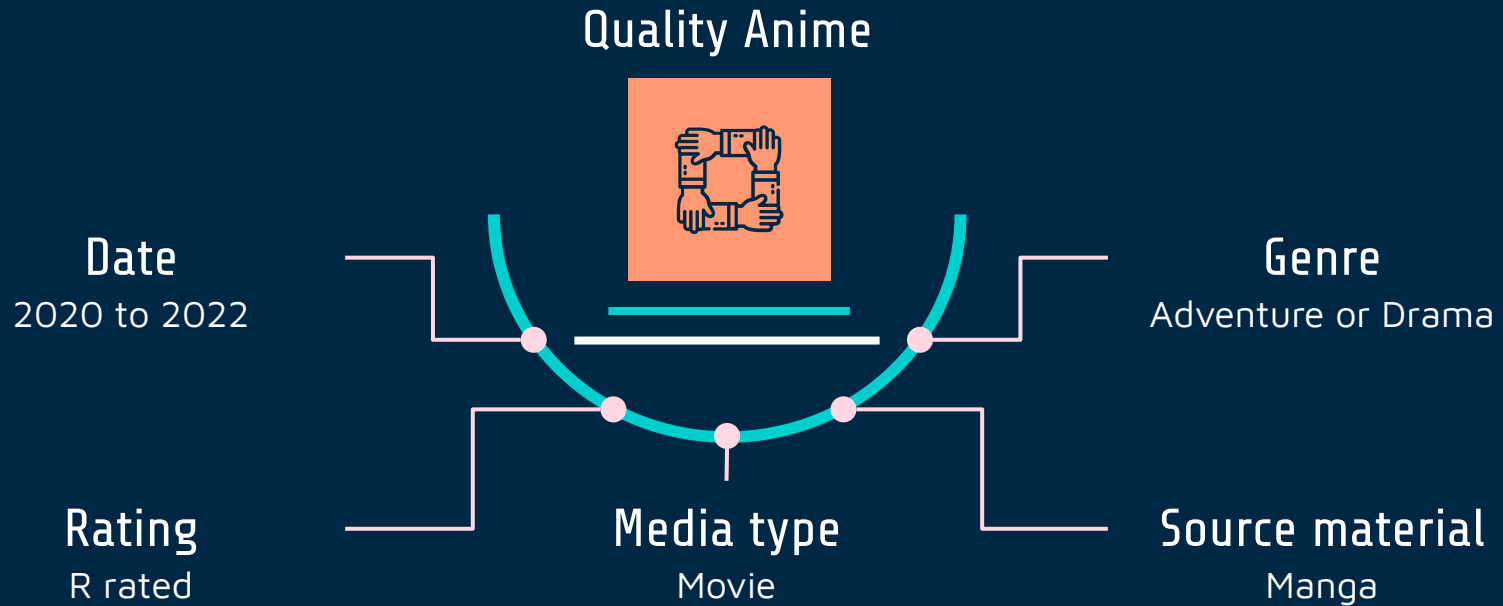
Important features using regression and classification: **Watch rate, start season year, rating, and start season year, season and popularity**

Important features using K means: **genre, media type and source material**

03

Things to note: Some of the Machine Learning models such as classification have certain flaws that may cause its results to hold less weight

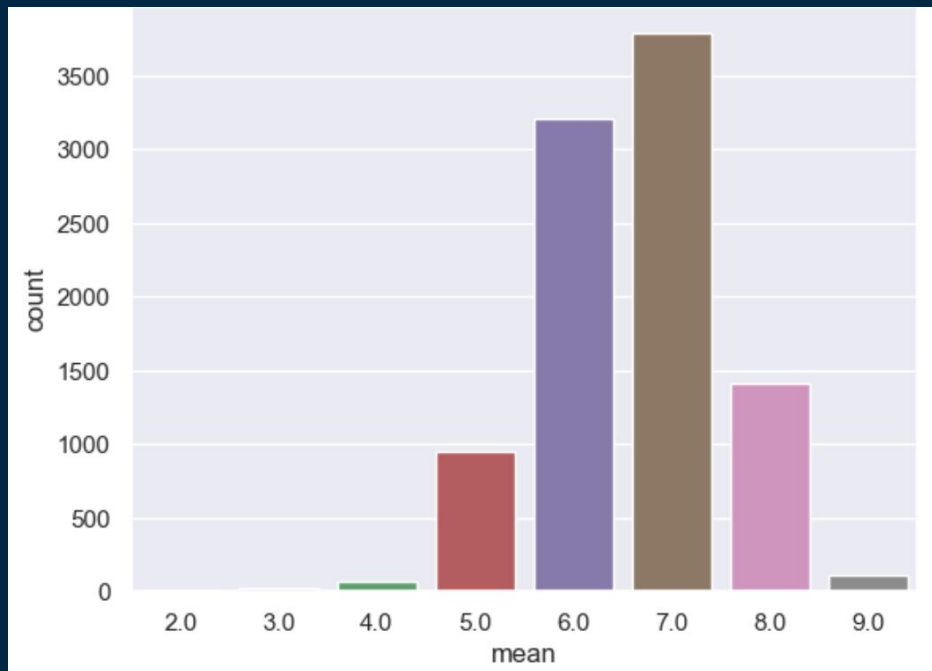
Conclusion



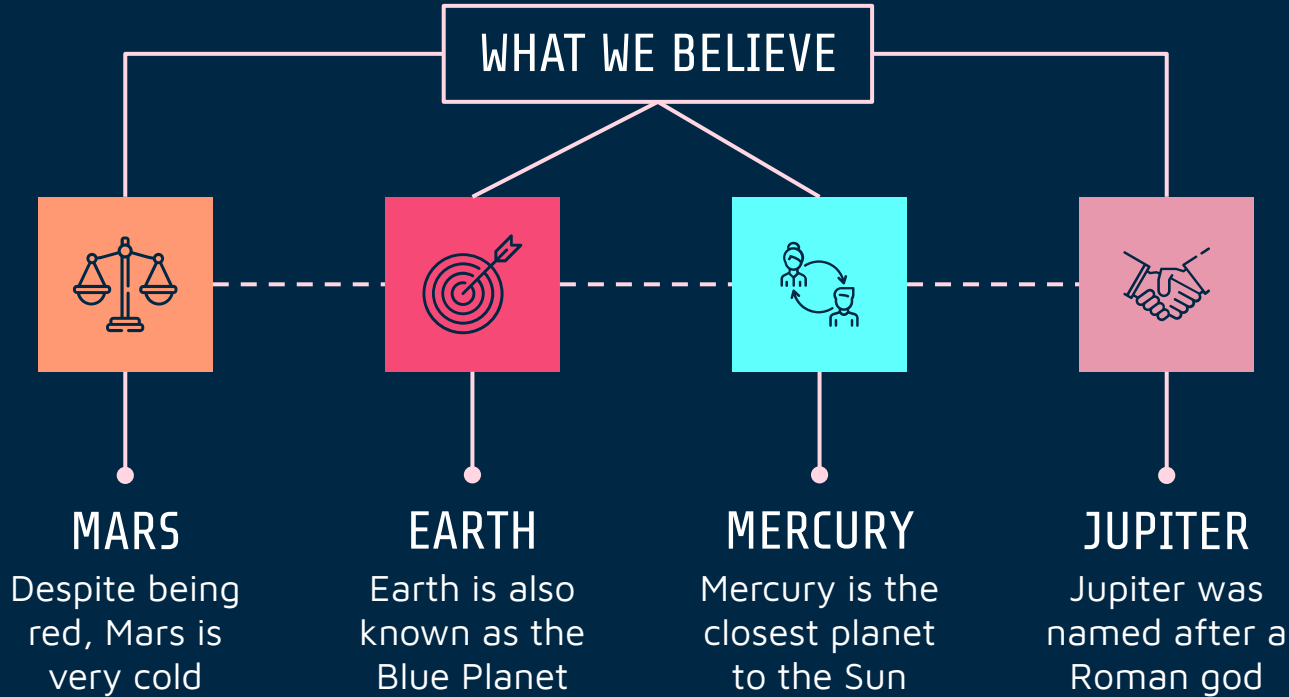
Conclusion

04

Limitations

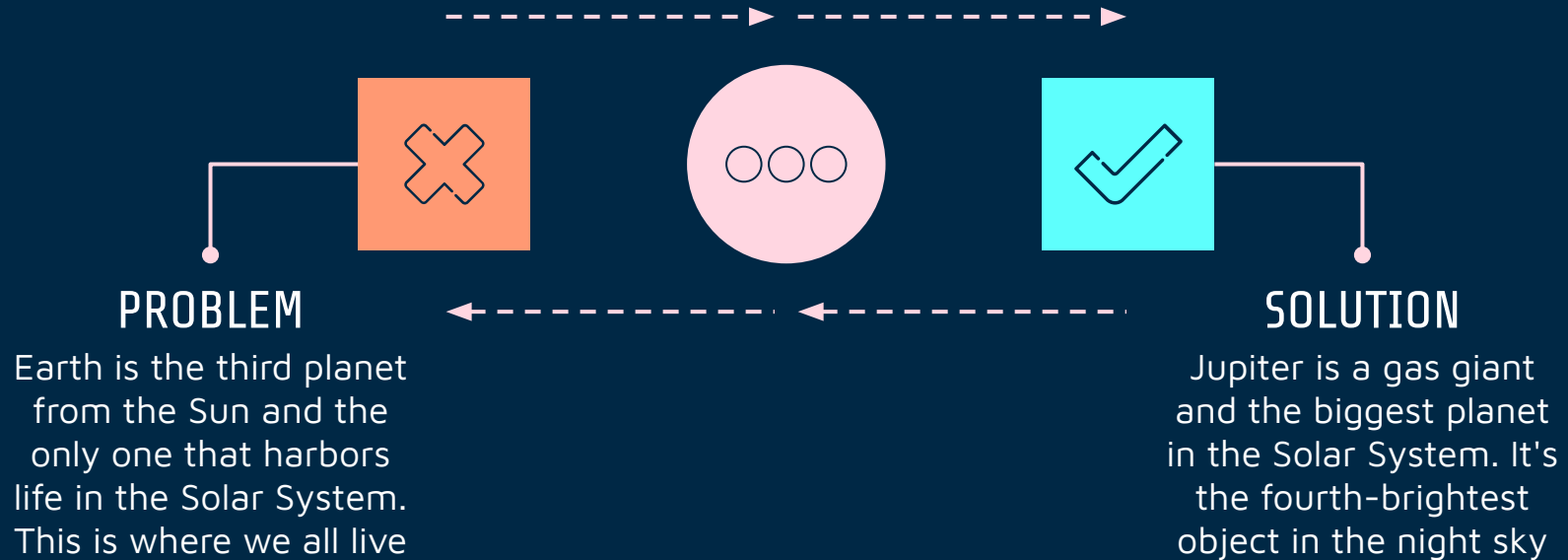


ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

OUR PROCESS

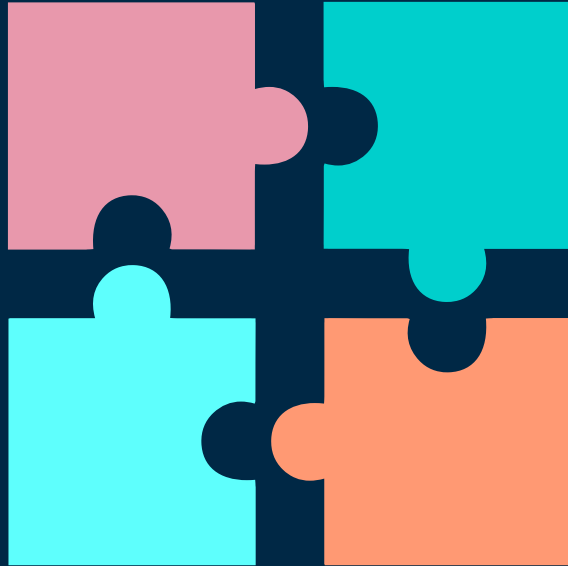


ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

THE PARTS



01

VENUS

Venus is the second planet from the Sun

02

MERCURY

Mercury is the closest planet to the Sun

03

JUPITER

Jupiter is the biggest planet of them all

04

MARS

Despite being red, Mars is very cold

ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

STEP BY STEP

JUPITER
Jupiter is the biggest
planet of them all

02



01

VENUS

Venus is the second
planet from the Sun

03

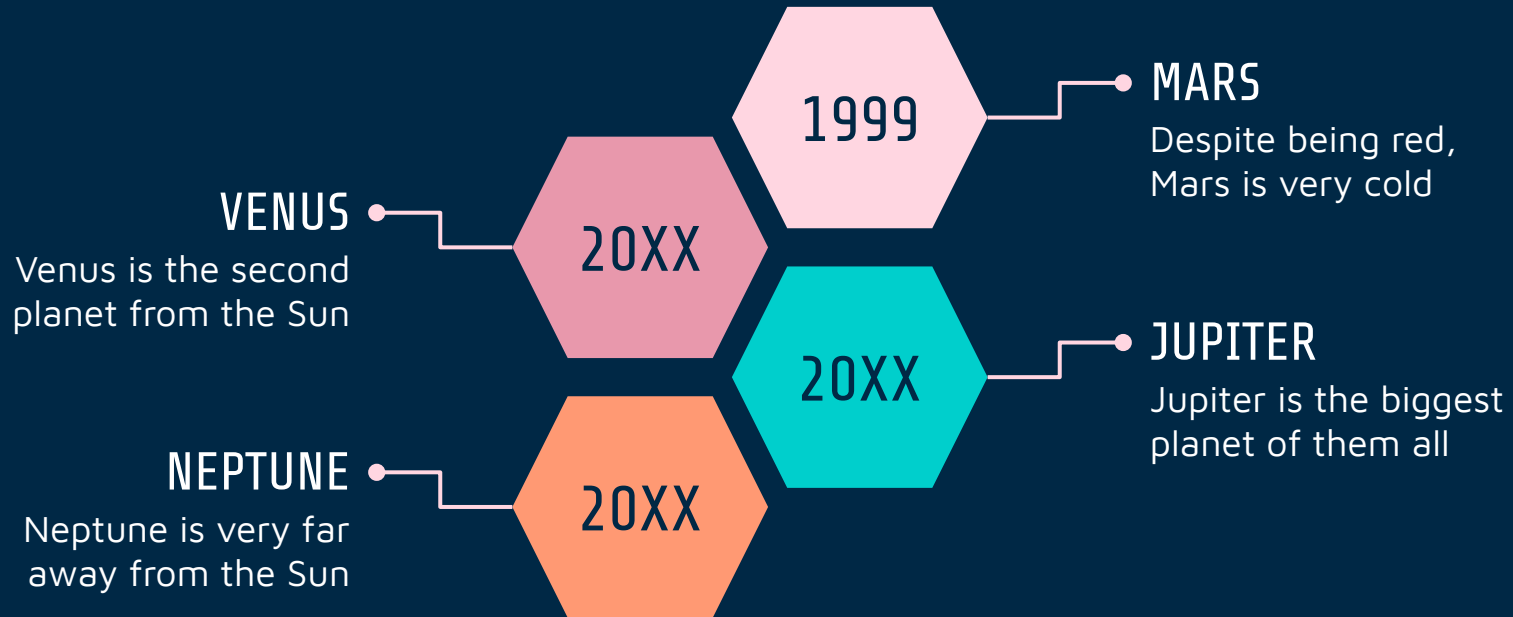
SATURN

Saturn is a gas giant
and has several rings



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

OUR EVOLUTION



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

01



DATA A

Despite being
red, Mars is
very cold

02



DATA B

Earth is also
known as the
Blue Planet

03



DATA C

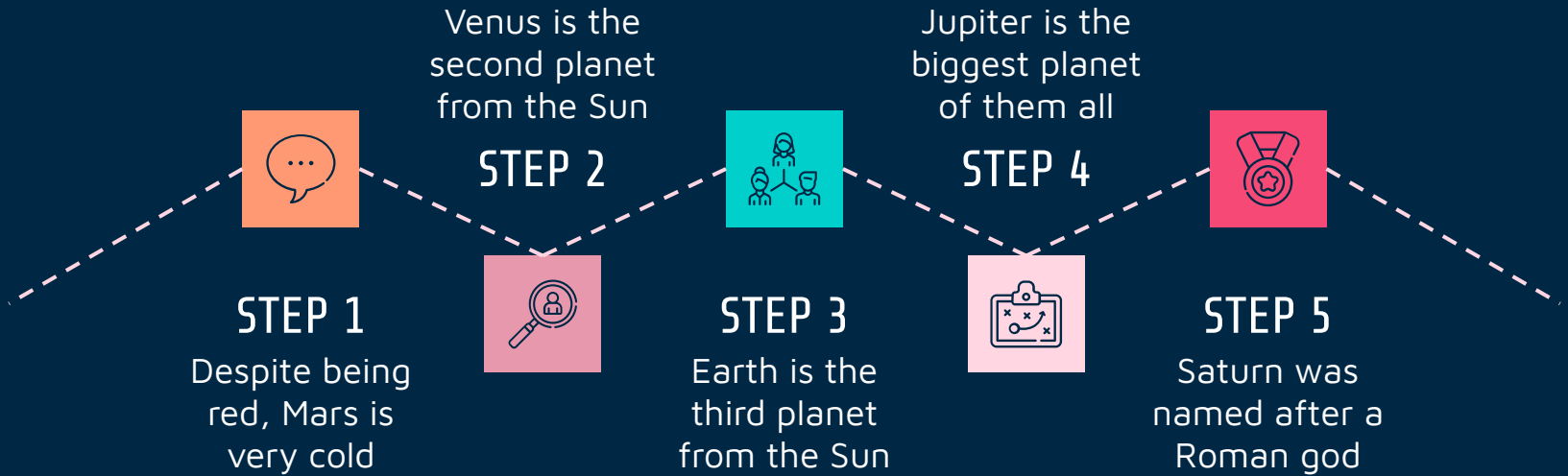
Jupiter was
named after a
Roman god

ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



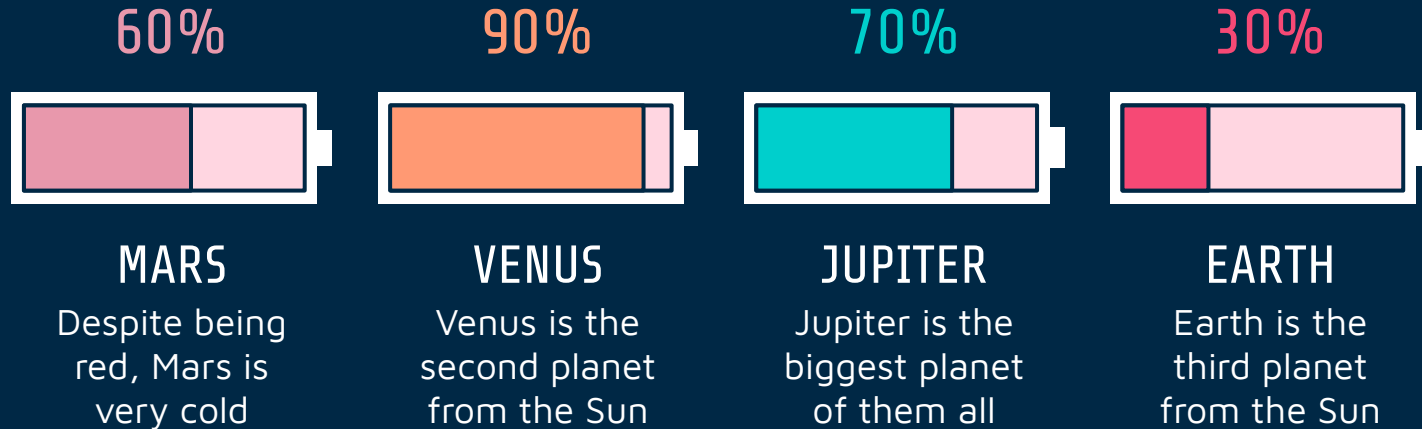
ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

SUCCESS ROAD



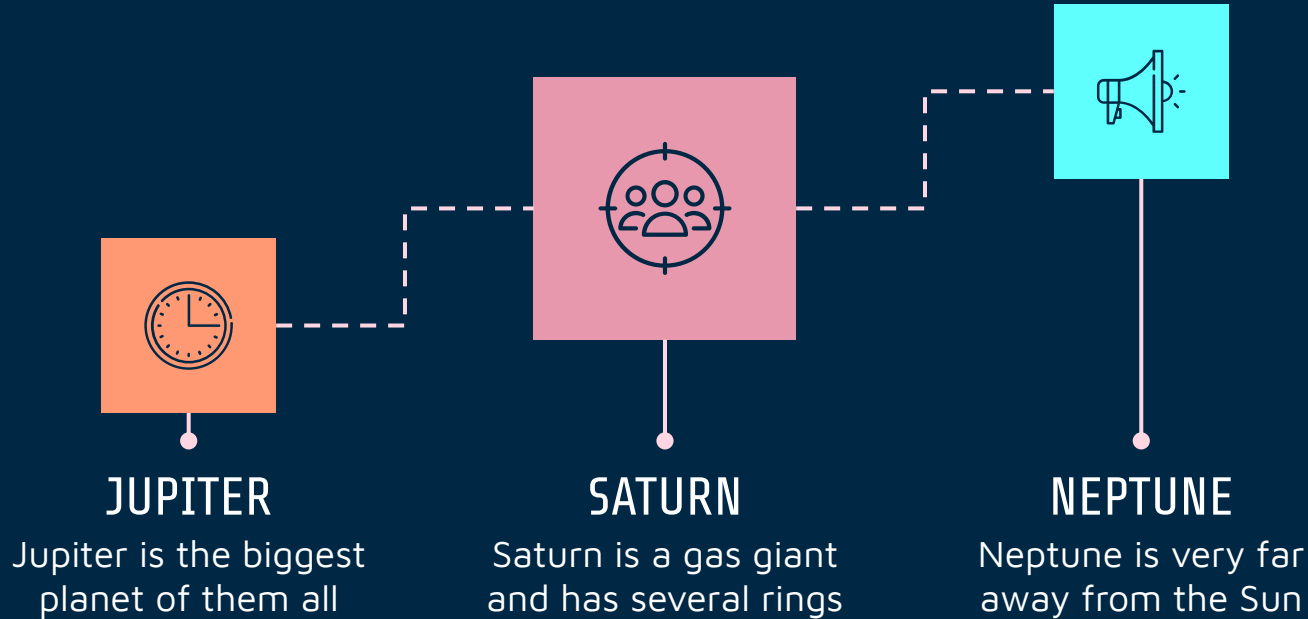
ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

OUR NUMBERS



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

OUR VALUES



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

WE STAND OUT

THEM

Earth is the planet with life



Jupiter is a huge gas giant



Neptune is far away from us



60%

40%

20%



100%

80%

60%

US

Mars is full of iron oxide dust



Saturn is the ringed planet



Mercury is a small planet



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

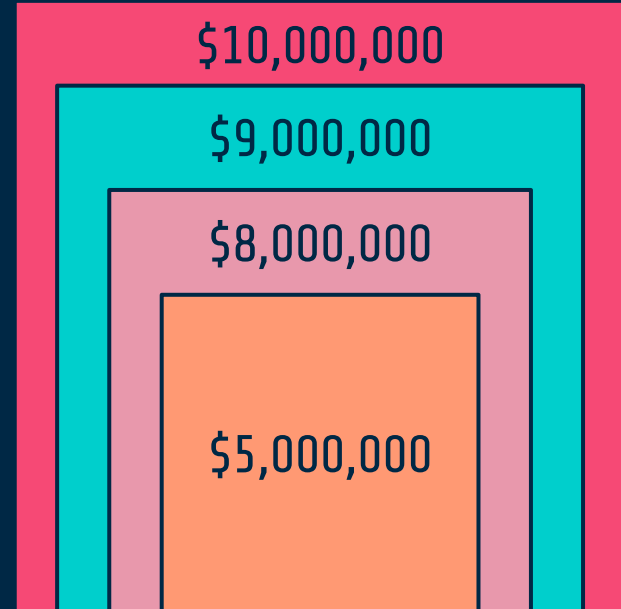
REVENUE FOR LAST YEAR


JANUARY
Mercury is the closest planet to the Sun


APRIL
Venus is the second planet from the Sun


AUGUST
Earth is the planet where we all live


DECEMBER
Saturn is a gas giant and has several rings

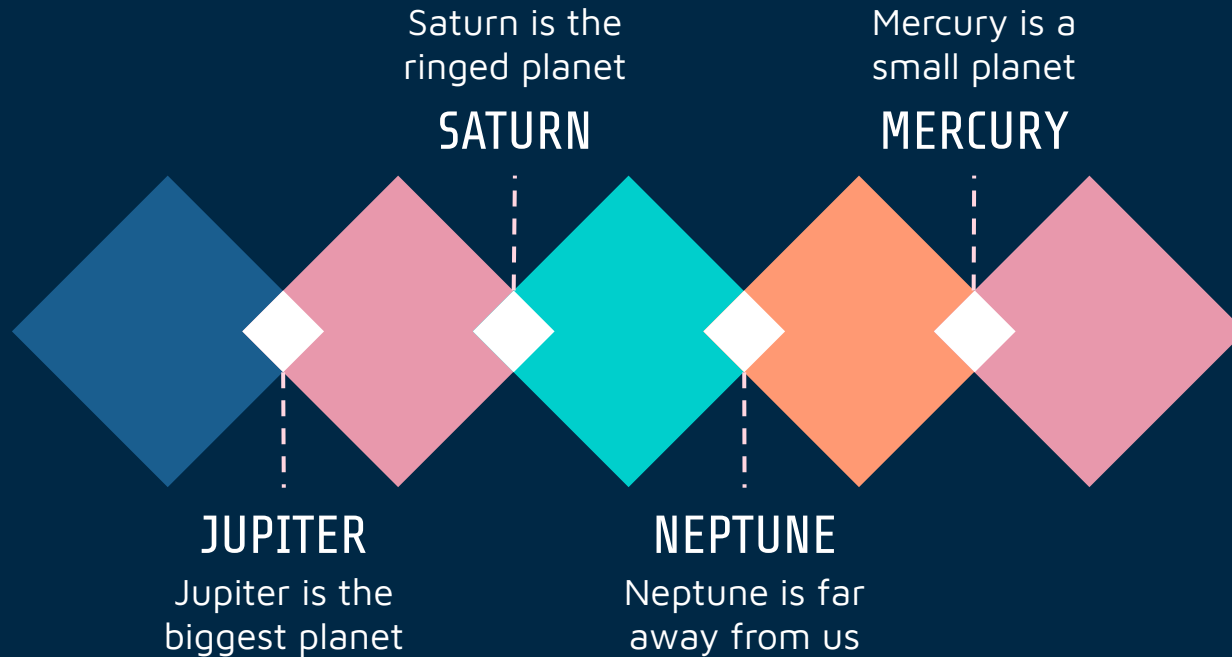


ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

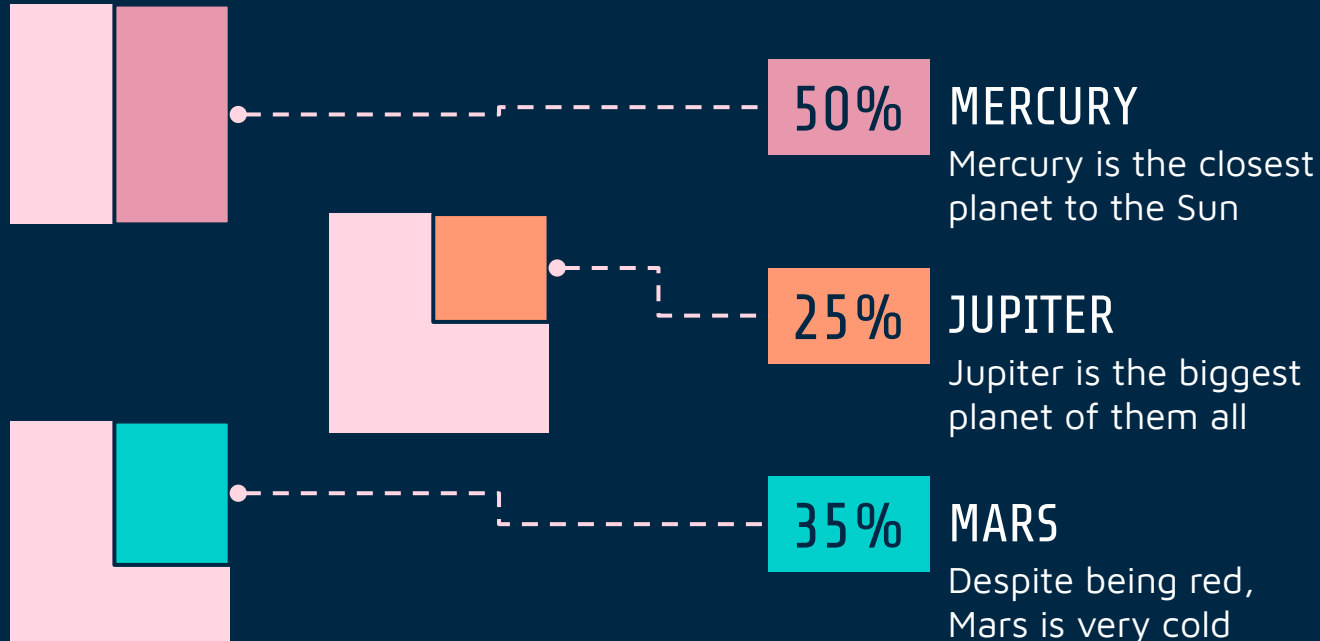
YEARLY PLANNING

JAN	Write down your own text	JUL	Write down your own text
FEB	Write down your own text	AGO	Write down your own text
MAR	Write down your own text	SEP	Write down your own text
APR	Write down your own text	OCT	Write down your own text
MAY	Write down your own text	NOV	Write down your own text
JUN	Write down your own text	DEC	Write down your own text

ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

OUR SERVICES



BASIC

- Mars is a red planet
- Despite being red, Mars is very cold

\$50

INTERMEDIATE

- Jupiter is a big planet
- Jupiter was named after a Roman god

\$100

PRO

- Saturn is a gas giant
- Saturn was named after a Roman god

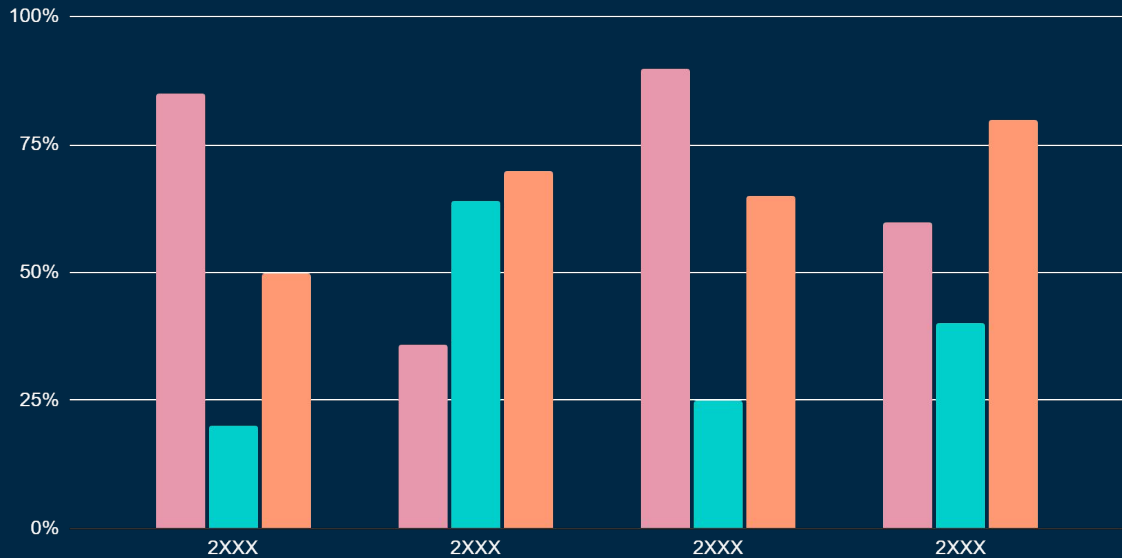
\$200

ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

INCOME
Saturn is the ringed planet

COST
Mars is full of iron oxide dust

PROFITS
Jupiter is the biggest planet



Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



VENUS

Venus is the second planet from the Sun

01

02

MARS

Despite being red, Mars is very cold

MERCURY

Mercury is the closest planet to the Sun

03

04

SATURN

Saturn is a gas giant and has several rings

JUPITER

Jupiter is the biggest planet of them all

05

06

NEPTUNE

Neptune is very far away from the Sun



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



SATURN
Saturn is the
ringed planet

NEPTUNE
Neptune is an
ice giant

45%

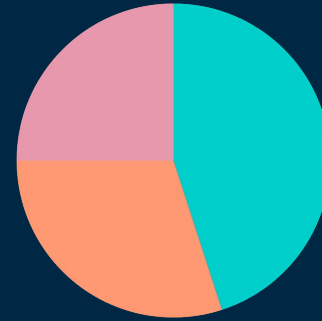
MALE

Venus is hot

25%

OTHER

Jupiter is big



30%

FEMALE

Mars is cold

Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

ANNUAL TASKS



TASK 1

Venus is hot



TASK 2

Earth has life

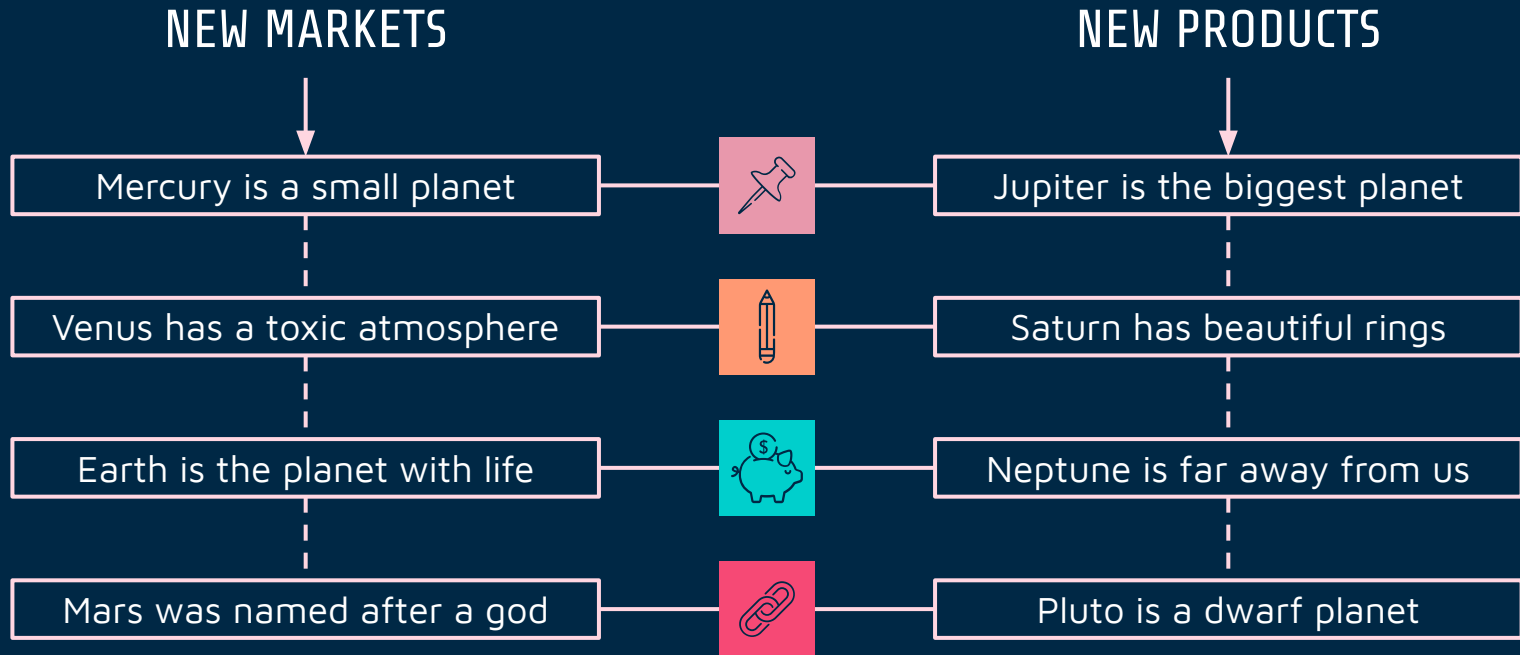


TASK 3

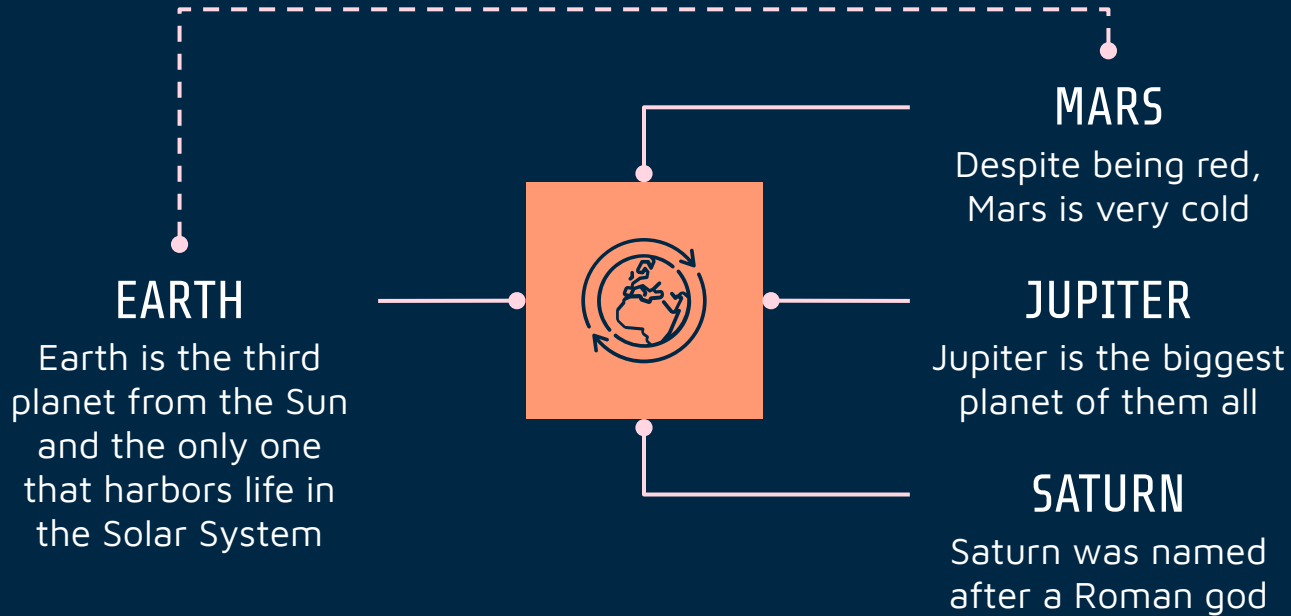
Mars is very cold

20XX	JAN-MAR	APR-JUN	JUL-SEP	OCT-DEC
TASK 1	Mercury is a very small planet			
TASK 2			Venus is hot	
TASK 3				Mars is cold

ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

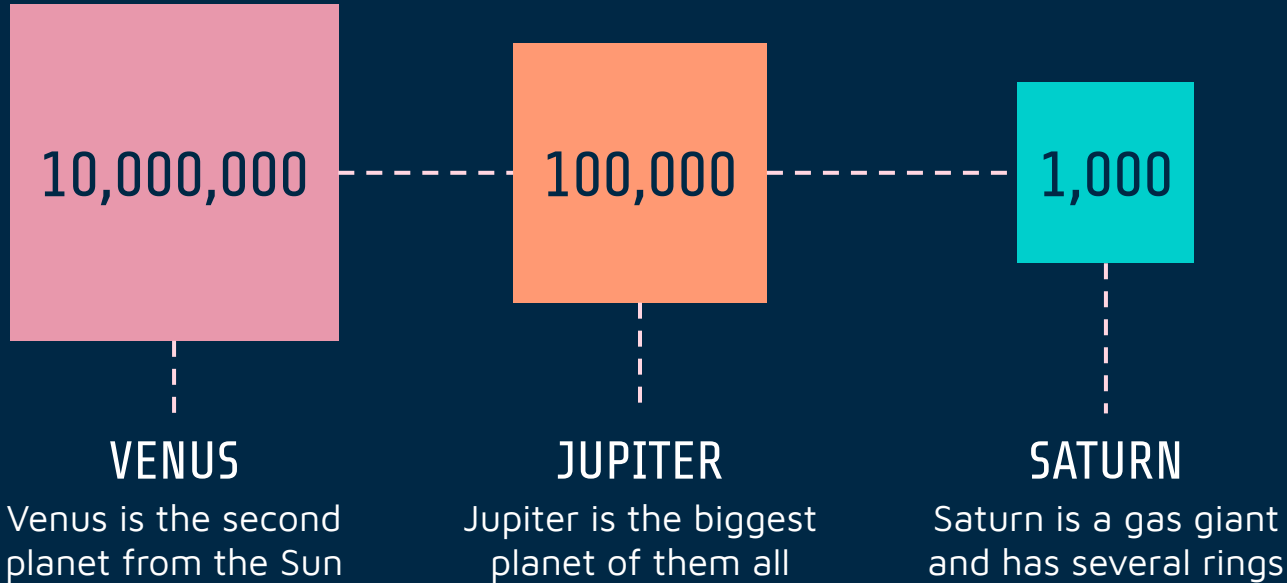


ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



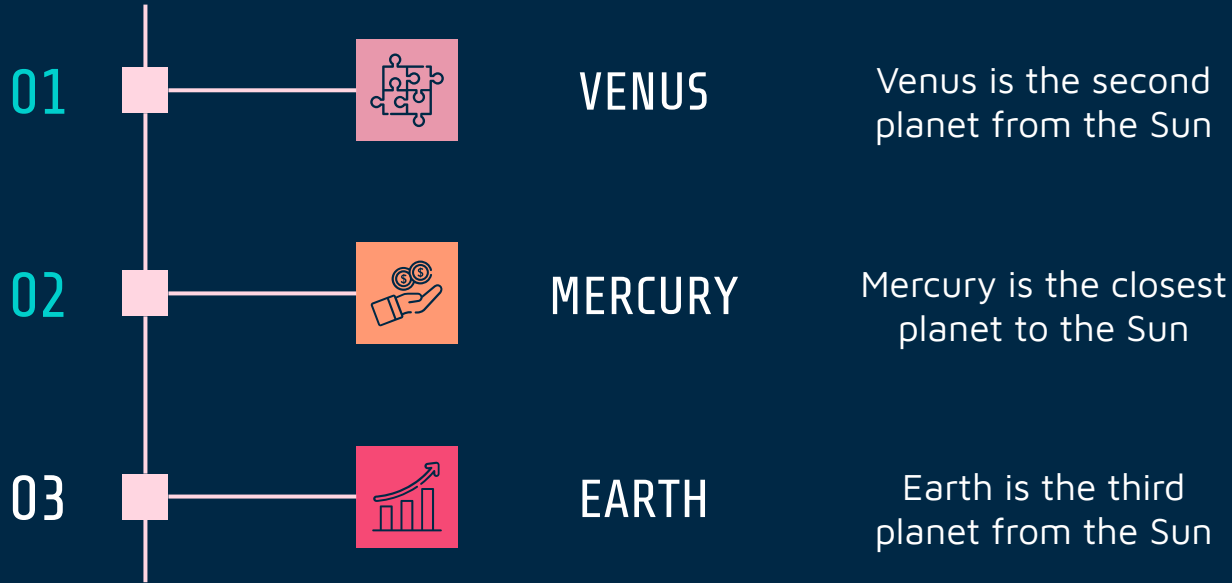
ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

MARKET SIZE



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS

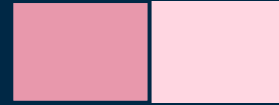
OUR TIMELINE



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



50%



Mercury is the closest planet to the Sun

30%



Venus is the second planet from the Sun

10%



Earth is the third planet from the Sun

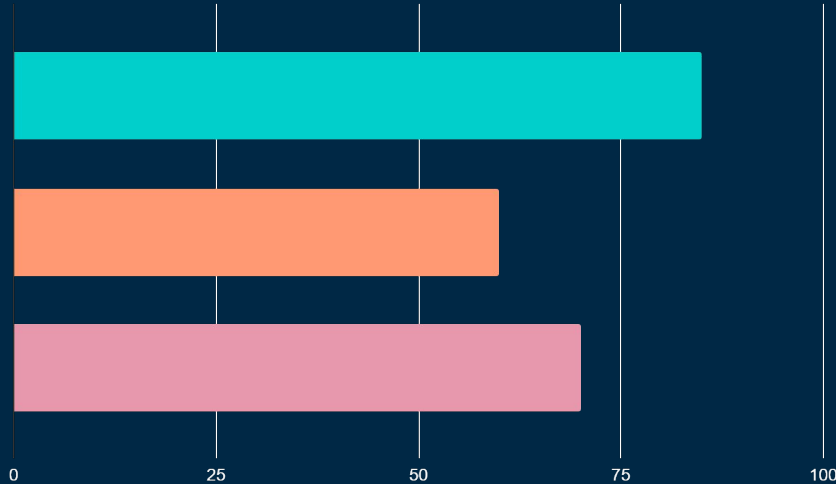
10%



Jupiter is the biggest planet of them all



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



85%

MANAGEMENT

Mercury is very small



60%

ANALYTICS

Mars is very cold



70%

BUSINESS

Venus is very hot



Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

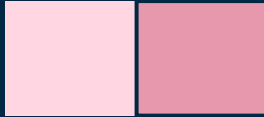


ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



NEPTUNE

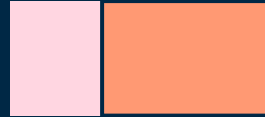
50%



JUPITER

Jupiter is the biggest planet of them all

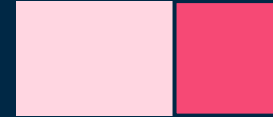
70%



MERCURY

Mercury is the closest planet to the Sun

40%



SATURN

Saturn was named after a Roman god



ABOUT DATA SCIENCE CONSULTING INFOGRAPHICS



TASKS FOR TODAY

- Venus is the second planet from the Sun
- Despite being red, Mars is very cold
- Saturn is not the only planet with rings

CHECKLIST

Add your task here	X
Add your task here	
Add your task here	
Add your task here	
Add your task here	X

Instructions for use

If you have a free account, in order to use this template, you must credit [Slidesgo](#) in your final presentation. Please refer to the next slide to read the instructions for premium users.

As a Free user, you are allowed to:

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

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute Slidesgo Content unless it has been expressly authorized by Slidesgo.
- Include Slidesgo Content in an online or offline database or file.
- Offer Slidesgo templates (or modified versions of Slidesgo templates) for download.
- Acquire the copyright of Slidesgo Content.

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

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

Instructions for use (premium users)

As a Premium user, you can use this template without attributing **Slidesgo**.

You are allowed to:

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

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

Infographics

You can add and edit some **infographics** to your presentation to present your data in a visual way.

- Choose your favourite infographic and insert it in your presentation using Ctrl C + Ctrl V or Cmd C + Cmd V in Mac.
- Select one of the parts and **ungroup** it by right-clicking and choosing “Ungroup”.
- **Change the color** by clicking on the paint bucket.
- Then **resize** the element by clicking and dragging one of the square-shaped points of its bounding box (the cursor should look like a double-headed arrow). Remember to hold Shift while dragging to keep the proportions.
- **Group** the elements again by selecting them, right-clicking and choosing “Group”.
- Repeat the steps above with the other parts and when you’re done editing, copy the end result and paste it into your presentation.
- Remember to choose the “**Keep source formatting**” option so that it keeps the design. For more info, please visit [our blog](#).

