FORMERLY Industries Helping Hands

THE iSCALE

Data Science | 30 Days of Machine Learning | Day - 18

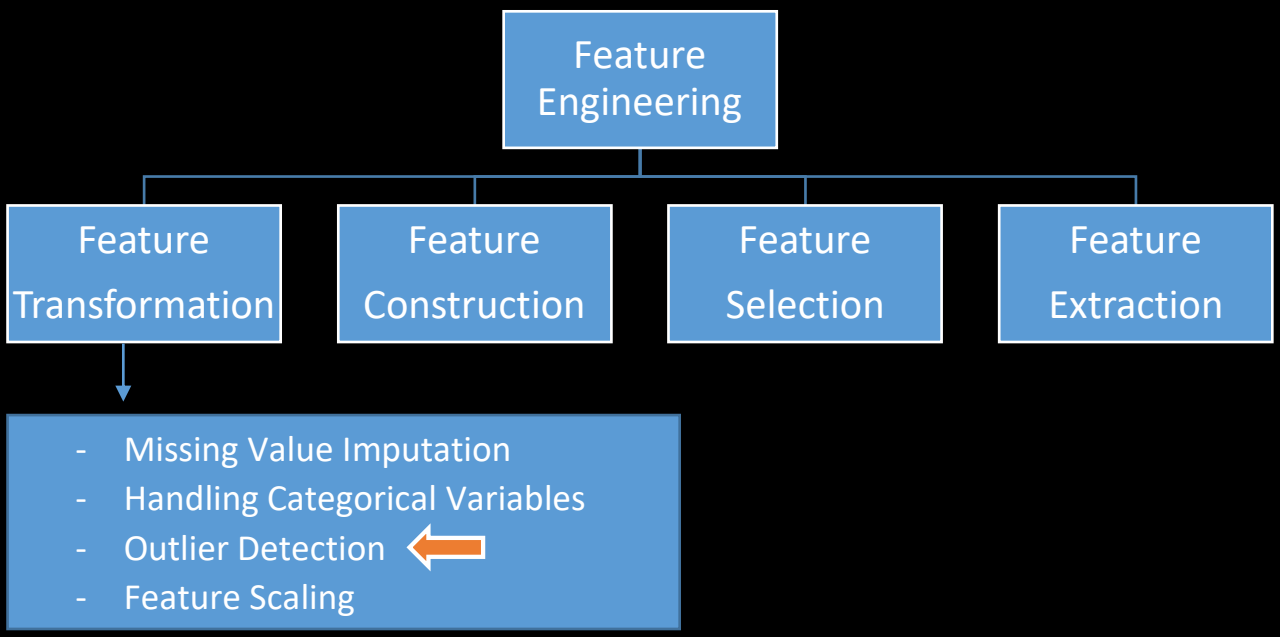
Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

----Today Topics | Day 18----

Outliers: Z score technique

- **Outliers removal using Z score treatment**
- **Z Score is applicable for normal distribution**
- **What is Standard deviation?**
- **Standard Normal Distribution (SND)**
- **Why are Z-Scores Important?**
- **How to Calculate "Z-Score"?**
- **What is 68 - 95- 99 Rule?**
- **Practice Problems For Z-Scores Calculation**

- Dataset Link GitHub: https://github.com/TheiScale/30_Days_Machine_Learning/



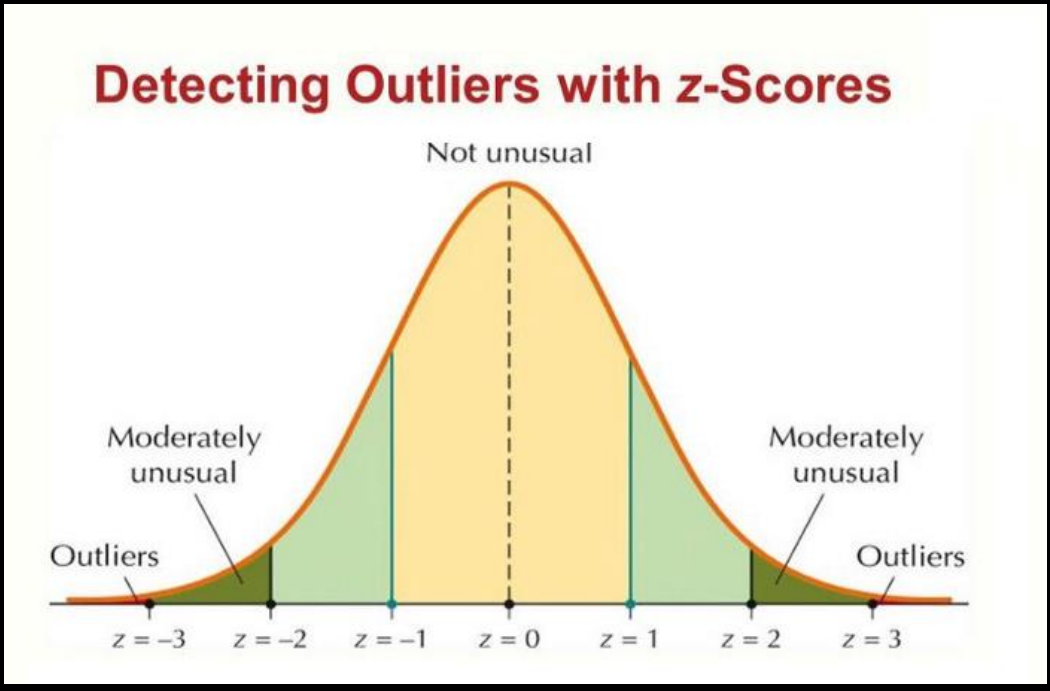
Techniques to detect & remove outliers: -

Z-score treatment: - This technique assumes that the column follows a normal distribution.

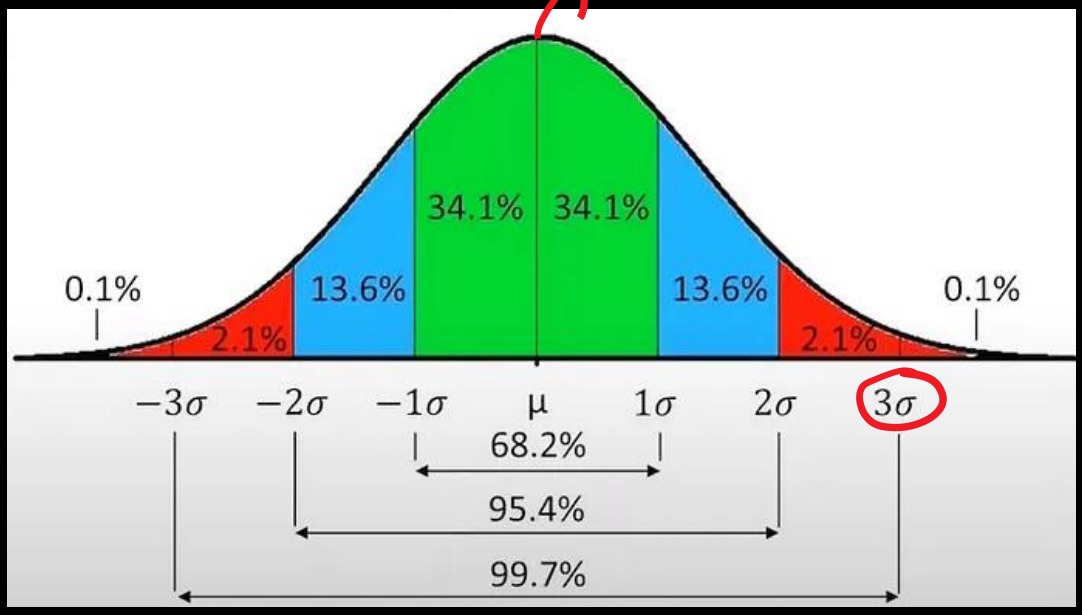
IQR (Interquartile Range) based filtering: - The IQR method involves calculating the range between the first quartile (Q1) and the third quartile (Q3).

Percentile method: - In this approach, a threshold is set based on percentiles. For example, if the threshold is set at 5%, any data point above the 95th percentile or below the 5th percentile is considered an outlier. These outliers can be removed or handled accordingly.

Winsorization:- Winsorization involves replacing outliers with values at a certain percentile, rather than removing them completely.



Outliers removal using Z score treatment:



μ + 1σ μ − 1σ

68.2.1.

μ + 2σ μ − 2σ

95.4.1.

μ + 3σ μ − 3σ

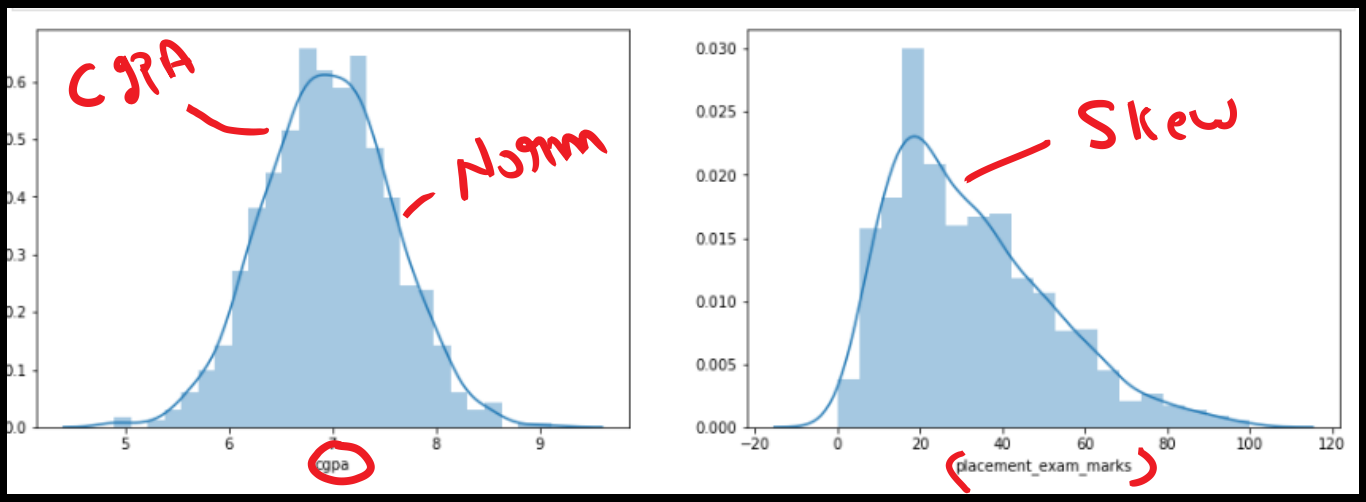
99.7.1.

Z Score is applicable for normal distribution:

https://www.scribbr.com/statistics/standard-normal-distribution/#:~:text=While%20data%20points%20are%20referred,is%20greater%20than%20the%20mean.

Z-score only applies to distributions derived from the normal distribution, or distributions that can be approximated by them. Many skewed distributions cannot. Therefore, much of the time, Z-score does not apply skewed distributions.

	cgpa	placement_exam_marks	placed
689	8.02	67.0	0
111	6.48	33.0	0
991	7.04	57.0	0
835	6.67	65.0	1
772	6.63	26.0	0

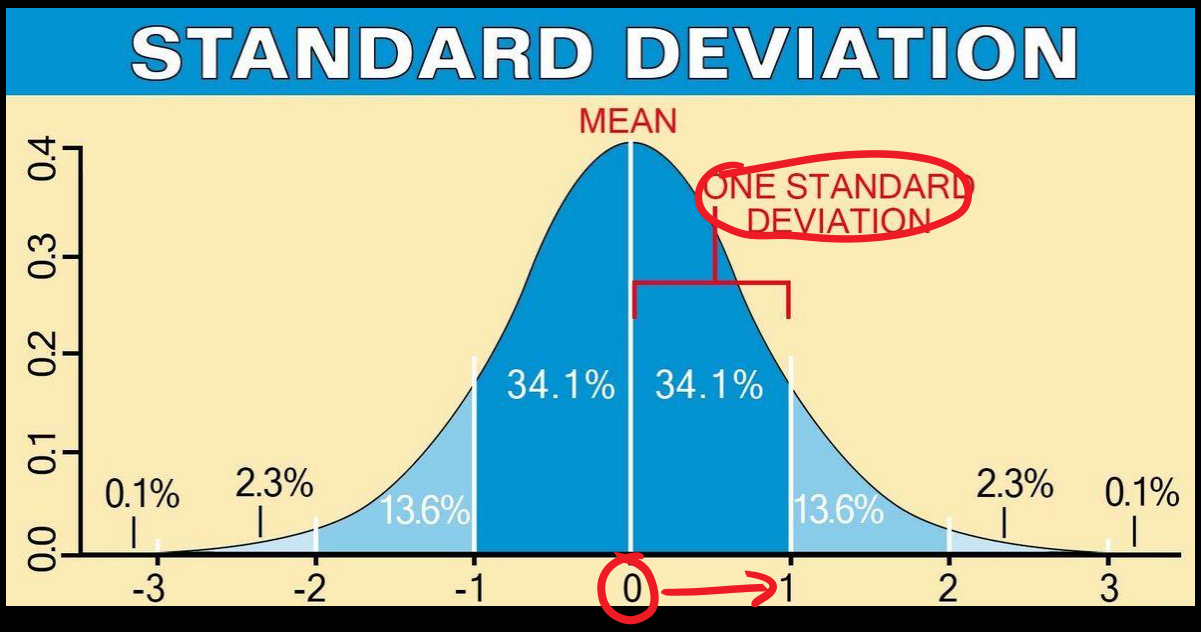


FORMERLY Industries Helping Hands

THE iSCALE

What is Standard deviation?

Standard Deviation is a measure which shows how much variation (such as spread) from the mean exists. The standard deviation indicates a “typical” deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set.



Standard Normal Distribution (SND)

1. The SND (i.e., z-distribution) is always the same shape as the raw score distribution. For example, if the distribution of raw scores is normally distributed, so is the distribution of z-scores.

2. The mean of any SND always = 0.

3. The standard deviation of any SND always = 1. Therefore, one standard deviation of the raw score (whatever raw value this is) converts into 1 z-score unit.

Blog: https://www.simplypsychology.org/z-score.html

Why are Z-Scores Important?

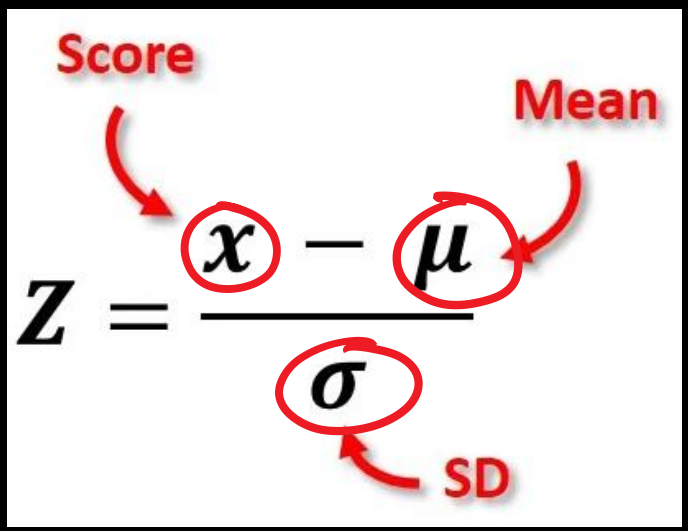
It is useful to standardize the values (raw scores) of a normal distribution by converting them into z-scores because:

1. It allows researchers to calculate the probability of a score occurring within a standard normal distribution;

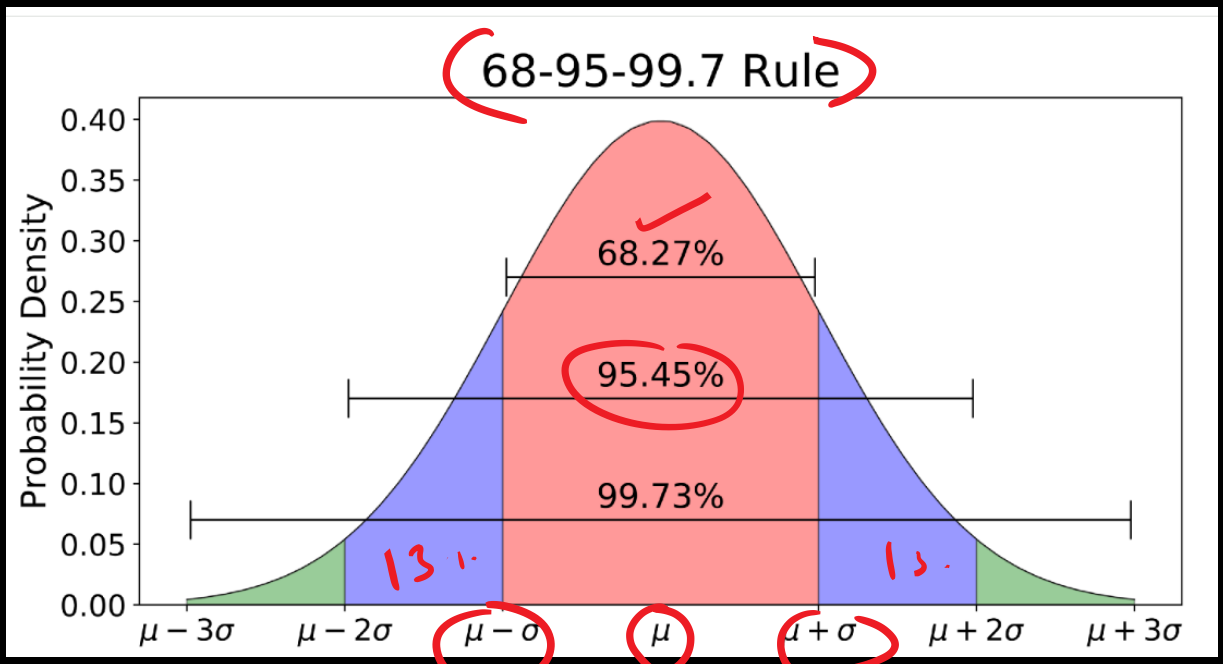
2. It enables us to compare two scores from different samples (which may have different means and standard deviations).

How to Calculate "Z-Score"?

The formula for calculating a z-score is z = (x-μ)/σ, where x is the raw score, μ is the population mean, and σ is the population standard deviation.

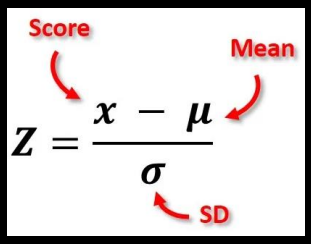


What is 68 - 95- 99 Rule:



Practice Problems For Z-Scores Calculation:

Problem 1: Scores on a psychological well-being scale range from 1 to 10, with an average score of 6 and a standard deviation of 2. What is the z-score for a person who scored 4?

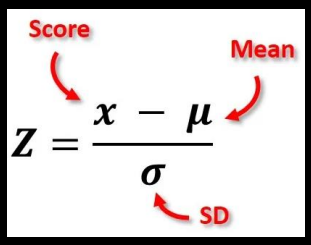


Solution 1:

Z-score = (4 – 6)/2 = -1

$$Z = \frac{x - \mu}{\sigma} = \frac{4 - 6}{2} = -2/2 = -1$$

Problem 2: On a measure of anxiety, a group of participants show a mean score of 35 with a standard deviation of 5. What is the z-score corresponding to a score of 30?

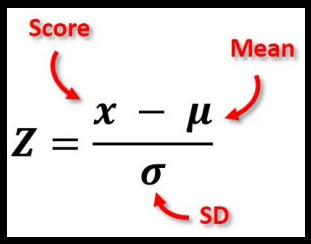


Solution 2:

Z-score = (30 – 35)/5 = -1

$$Z = \frac{x - \mu}{\sigma} = \frac{30 - 35}{5} = \frac{-5}{5} = -1$$

Problem 3: In a study on sleep, participants report an average of 7 hours of sleep per night, with a standard deviation of 1 hour. What is the z-score for a person reporting 5 hours of sleep?

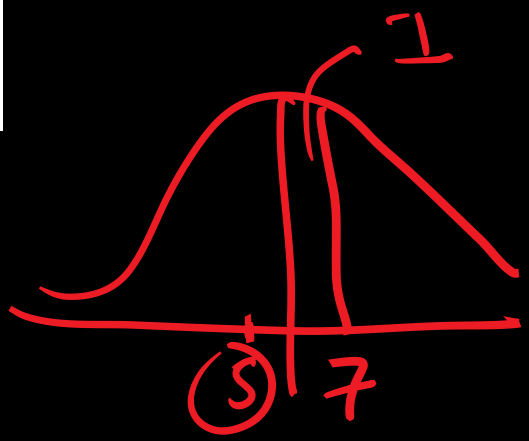


Solution 3:

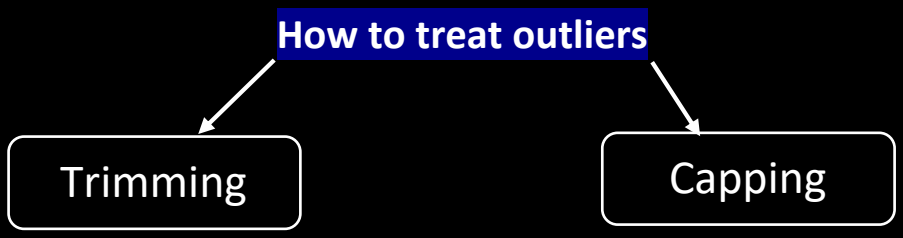
Z-score = (5 – 7)/1 = -2

$$Z = \frac{x - \mu}{\sigma} = \frac{5 - 7}{1}$$

$$Z = -2$$



Step 1: Finding the boundary values (Highest & Lowest)
Step 2: Finding the outliers
Step 3: Treat outliers with suitable technique.



<Start Coding>

```
#Import Library

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#Import Dataset

df = pd.read_csv('placement.csv')

----
df

----
df.shape

----
df.sample(5)
```

```
#Plot Show in CGPA and Placement Marks

plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.distplot(df['cgpa'])

plt.subplot(1,2,2)
sns.distplot(df['placement_exam_marks'])

plt.show()

#Print Mean | Std | Min & Max Value

print("Mean value of cgpa",df['cgpa'].mean())
print("Std value of cgpa",df['cgpa'].std())
print("Min value of cgpa",df['cgpa'].min())
print("Max value of cgpa",df['cgpa'].max())


#Approach 1

#Step 1:Finding the boundary values (Highest & Lowest)

print("Highest allowed",df['cgpa'].mean() + 3*df['cgpa'].std())

print("Lowest allowed",df['cgpa'].mean() - 3*df['cgpa'].std())

#Step 2: Finding the outliers

df[(df['cgpa'] > 8.80) | (df['cgpa'] < 5.11)]
```

```
#Step 3: Treat Outliers with Trimming

new_df = df[(df['cgpa'] < 8.80) & (df['cgpa'] >
5.11)]
new_df

# Approach 2 : Calculating the Zscore

df['cgpa_zscore'] = (df['cgpa'] -
df['cgpa'].mean())/df['cgpa'].std()

----
df.head()

#CGPA Score More then 3

df[df['cgpa_zscore'] > 3]

#CGPA Score Less then 3

df[df['cgpa_zscore'] < -3]

#Show or Merge Both CGPA

df[(df['cgpa_zscore'] > 3) | (df['cgpa_zscore'] < -
3)]


# Apply Trimming

new_df = df[(df['cgpa_zscore'] < 3) &
(df['cgpa_zscore'] > -3)]

----

new_df
```

FORMERLY Industries Helping Hands

THE iSCALE

Day 18: Curious Data Minds

Art of Storytelling in Data Science:

- Why Story telling is important for "Data Science" Interview?

Storytelling is important for data science interviews because:

1. **Communication**: It helps explain complex data findings to non-technical people.
2. **Context**: It puts data analysis into a meaningful story, explaining why it's important.
3. **Engagement**: Stories capture attention and persuade people better than just data.
4. **Simplicity**: It simplifies complex data concepts, making them easier to understand.
5. **Memorability**: Stories are more memorable, ensuring that insights stick with people.

In interviews, storytelling shows you can explain data clearly, understand its context, and influence decisions—valuable skills in data science.

https://eightify.app/summary/miscellaneous/why-jeff-bezos-banned-powerpoint-at-amazon-lex-fridman-podcast-clips

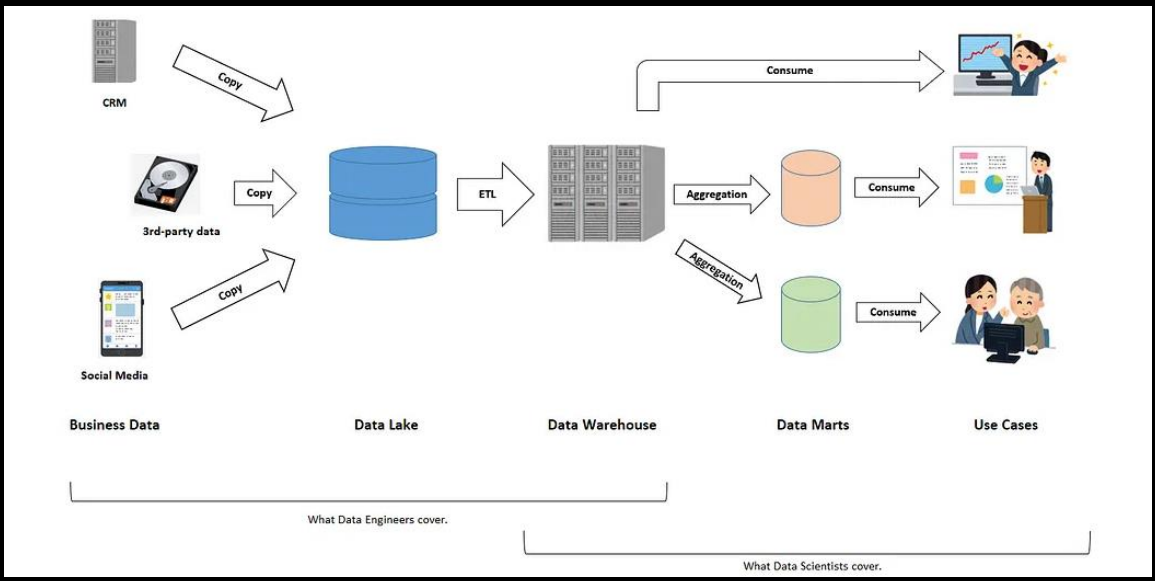
- Focus point to remember in Interview

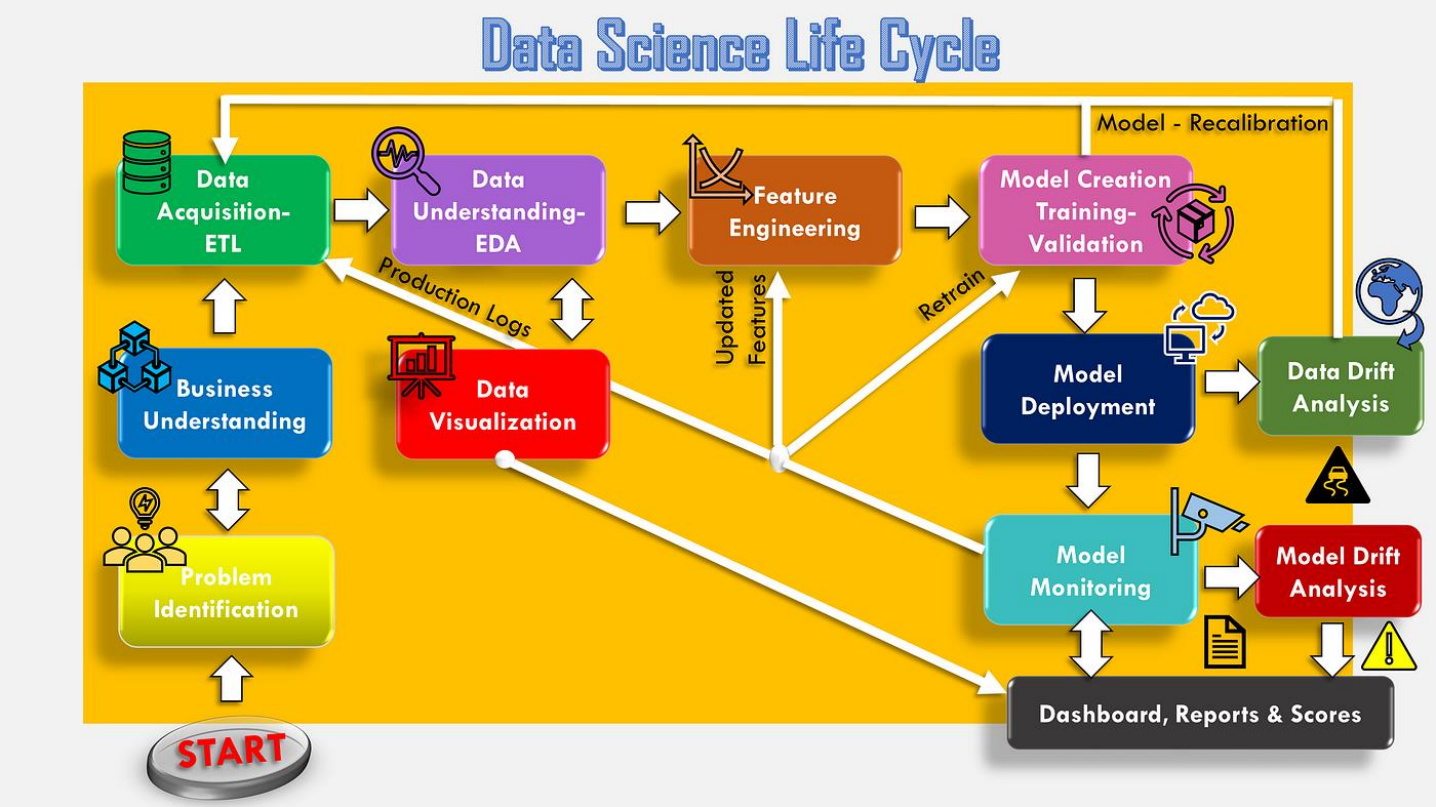
- **Tell interviewers what you know?**
- **Focus on Data Storytelling (About Projects)**

- How should be start the Data Story telling?

1. **Introduction**: Begin by introducing your project and why it's important.
2. **Hook**: Grab your interviewer's attention with an interesting fact or story related to your data.
3. **Problem Statement**: Clearly state the problem you're trying to solve or the question you're exploring.
4. **Objective**: Explain what you aim to achieve through your analysis.
5. **Preview of Insights**: Give a quick overview of the main findings or insights you'll be sharing.

- "Tell about your projects."
- "Discuss the architecture used."
- "Focus on the Life Cycle of Data Science Projects."
- "Focus on your roles in the team."
- "Tell about the challenges you faced and how you handled them."
  https://towardsdatascience.com/fundamentals-of-data-architecture-to-help-data-scientists-understand-architectural-diagrams-better-7bd26de41c66




Data Science Life Cycle
Model - Recalibration
Data Acquisition- ETL
Data Understanding- EDA
Feature Engineering
Model Creation Training- Validation
Production Logs
Updated Features
Retrain
Business Understanding
Data Visualization
Model Deployment
Data Drift Analysis
Problem Identification
Model Monitoring
Model Drift Analysis
START
Dashboard, Reports & Scores