MySkill IdScore PEFINDO Credit Bureau

Portofolio - Short Class

Data Science Introduction

Owner: Muhammad Aziz Habiburrahim

Build your skill and portfolio via myskill.id/bootcamp

Course Summary

Silahkan dapat menuliskan summary dari kelas yang diikuti hari ini.

Poin Belajar	Rangkuman
Introduction to Data Science	Data Science adalah bidang yang menggabungkan statistik, pemrograman, dan analisis data untuk menemukan wawasan yang berharga. Proses dalam Data Science mencakup pengumpulan, pembersihan, eksplorasi, analisis, dan visualisasi data. Data Scientist menggunakan berbagai teknik seperti machine learning dan algoritma statistik untuk membuat prediksi atau keputusan berbasis data. Exploratory Data Analysis (EDA) menjadi tahap penting dalam memahami pola dan distribusi data sebelum membangun model. Selain itu, machine learning memungkinkan sistem untuk belajar dari data tanpa pemrograman eksplisit. Model prediktif digunakan dalam berbagai bidang seperti keuangan, kesehatan, dan pemasaran untuk mengoptimalkan strategi bisnis. Data Science juga memanfaatkan big data dan cloud computing untuk mengelola dan memproses data dalam skala besar. Pemahaman tentang data cleaning sangat penting untuk memastikan kualitas data yang baik sebelum dianalisis. Visualisasi data membantu menyampaikan temuan secara lebih jelas kepada pengambil keputusan. Secara keseluruhan, Data Science berperan penting dalam mengubah data mentah menjadi informasi yang bernilai bagi berbagai industri.

Course Summary

Silahkan dapat menuliskan summary dari kelas yang diikuti hari ini.

Poin Belajar	Rangkuman
The Workflow of Data Science	The Workflow of Data Science mencakup beberapa tahapan utama dalam mengolah data untuk mendapatkan wawasan berharga. Tahap pertama adalah **pengumpulan data**, yang bisa berasal dari berbagai sumber seperti database, API, atau sensor. Selanjutnya, dilakukan **pembersihan data (data cleaning)** untuk menangani data yang hilang, duplikat, atau tidak valid. Setelah itu, tahap **Exploratory Data Analysis (EDA)** dilakukan untuk memahami pola, distribusi, dan hubungan antar variabel. Kemudian, **pemrosesan fitur (feature engineering)** dilakukan untuk memilih atau membuat variabel yang lebih relevan bagi model. Tahap berikutnya adalah **pemilihan dan pelatihan model machine learning** dengan algoritma yang sesuai dengan tujuan analisis. Setelah model dilatih, dilakukan **evaluasi model** menggunakan metrik seperti akurasi, precision, dan recall. Jika hasilnya belum optimal, dilakukan **penyetelan parameter (hyperparameter tuning)** untuk meningkatkan performa model. Setelah model siap, tahap **deploy model** dilakukan agar bisa digunakan dalam lingkungan produksi. Terakhir, model yang sudah diterapkan harus **dimonitor dan diperbarui** secara berkala untuk menjaga akurasinya sesuai dengan perubahan data baru.

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[94] # import library
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     from google.colab import drive
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[95] # read and load dataset
     df = pd.read_csv('/content/drive/MyDrive/Course/Datasets/titanic-case.csv')
     df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[78] # rename variable
     df = df.rename(columns={'PassengerId': 'passenger_id', 'Survived': 'survived', 'Pclass' : 'p_class', 'Name' : 'name',
                             'Sex': 'sex', 'Age': 'age', 'SibSp' : 'sibsp', 'Parch' : 'parch',
                             'Ticket': 'ticket', 'Fare': 'fare', 'Cabin' : 'cabin', 'Embarked' : 'embarked'})
     df.head()
```

	passenger_id	survived	p_class	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[79] # summary statistics dataset
     df.describe()
```

	passenger_id	survived	p_class	age	sibsp	parch	fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[80] # checking for missing values
     df.isnull().sum()
```

	0
passenger_id	0
survived	0
p_class	0
name	0
sex	0
age	177
sibsp	0
parch	0
ticket	0
fare	0
cabin	687
embarked	2

dtype: int64

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[81] # handling missing values in 'age' by filling with the median age
     df['age'].fillna(df['age'].median(), inplace=True)
     # handling missing values in 'embarked' by filling with the mode
     df['embarked'].fillna(df['embarked'].mode()[0], inplace=True)
     # handling missing values in 'fare' by filling with the median fare
     df['fare'].fillna(df['fare'].median(), inplace=True)
     # dropping the 'cabin' column due to a high number of missing values
     df.drop(columns=['cabin'], inplace=True)
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[82] # checking for missing values again
     missing_values_after = df.isnull().sum()
     print("\nMissing values in the dataset after handling:")
     print(missing_values_after)
```

```
Missing values in the dataset after handling:
passenger_id    0
survived        0
p_class         0
name            0
sex             0
age             0
sibsp           0
parch           0
ticket          0
fare            0
embarked        0
dtype: int64
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[83] # checking for duplicates
     duplicates = df.duplicated().sum()
     print("\nNumber of duplicates in the dataset:")
     print(duplicates)
```

```
Number of duplicates in the dataset:
0
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[84] # dropping duplicates
     df.drop_duplicates(inplace=True)
```

```
[85] # checking for duplicates again
     duplicates_after = df.duplicated().sum()
     print("\nNumber of duplicates in the dataset after removing duplicates:")
     print(duplicates_after)
```

```
Number of duplicates in the dataset after removing duplicates:
0
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[86] # displaying the first few rows dataset
     print("\nFirst few rows of the dataset after handling missing values and duplicates:")
     print(df.head())
```

```
First few rows of the dataset after handling missing values and duplicates:
   passenger_id  survived  p_class  \
0             1         0        3
1             2         1        1
2             3         1        3
3             4         1        1
4             5         0        3

                                                name     sex   age  sibsp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0

   parch            ticket     fare embarked
0      0         A/5 21171   7.2500        S
1      0          PC 17599  71.2833        C
2      0  STON/O2. 3101282   7.9250        S
3      0            113803  53.1000        S
4      0            373450   8.0500        S
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[87] # summary statistics dataset
     print("\nSummary statistics of the dataset:")
     print(df.describe())
```

```
Summary statistics of the dataset:
       passenger_id    survived     p_class         age       sibsp  \
count    891.000000  891.000000  891.000000  891.000000  891.000000
mean     446.000000    0.383838    2.308642   29.361582    0.523008
std      257.353842    0.486592    0.836071   13.019697    1.102743
min        1.000000    0.000000    1.000000    0.420000    0.000000
25%      223.500000    0.000000    2.000000   22.000000    0.000000
50%      446.000000    0.000000    3.000000   28.000000    0.000000
75%      668.500000    1.000000    3.000000   35.000000    1.000000
max      891.000000    1.000000    3.000000   80.000000    8.000000

             parch        fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

Owner : Muhammad Aziz Habiburrahim

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[88] # information about the dataset
     print("\nInformation about the dataset:")
     print(df.info())
```

```
Information about the dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   passenger_id  891 non-null    int64
 1   survived      891 non-null    int64
 2   p_class       891 non-null    int64
 3   name          891 non-null    object
 4   sex           891 non-null    object
 5   age           891 non-null    float64
 6   sibsp         891 non-null    int64
 7   parch         891 non-null    int64
 8   ticket        891 non-null    object
 9   fare          891 non-null    float64
 10  embarked      891 non-null    object
dtypes: float64(2), int64(5), object(4)
memory usage: 76.7+ KB
None
```

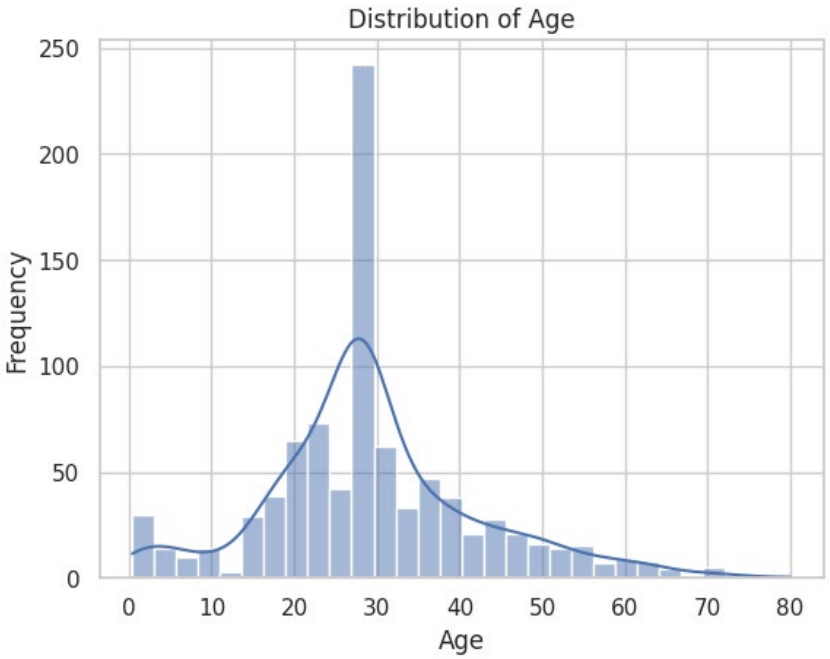
Practice using Google Colab

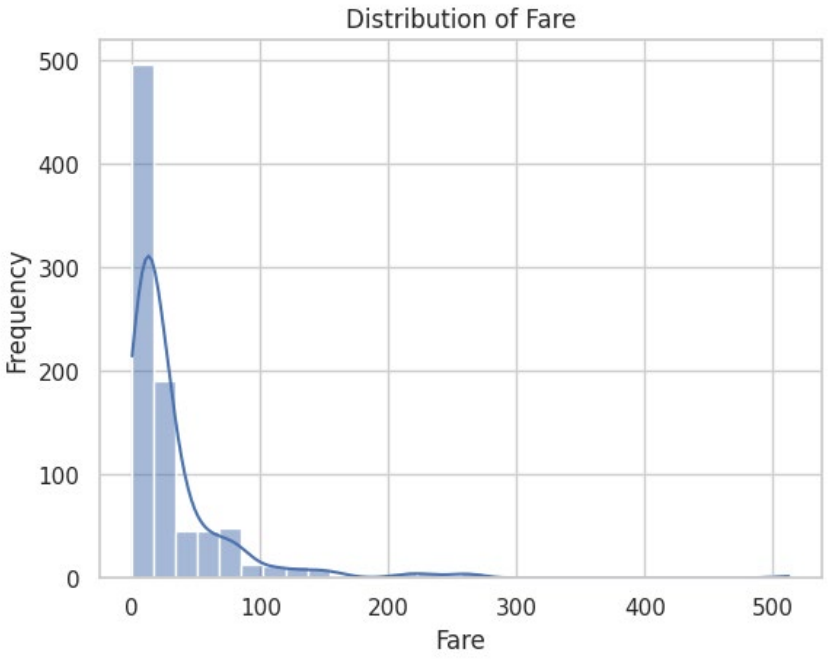
Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[89] # visualizing the distribution of age and fare
     sns.histplot(df['age'].dropna(), bins=30, kde=True)
     plt.title('Distribution of Age')
     plt.xlabel('Age')
     plt.ylabel('Frequency')
     plt.show()
     sns.histplot(df['fare'].dropna(), bins=30, kde=True)
     plt.title('Distribution of Fare')
     plt.xlabel('Fare')
     plt.ylabel('Frequency')
     plt.show()
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.





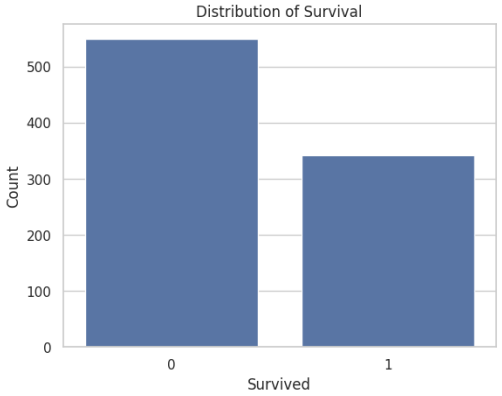
Owner : Muhammad Aziz Habiburrahim

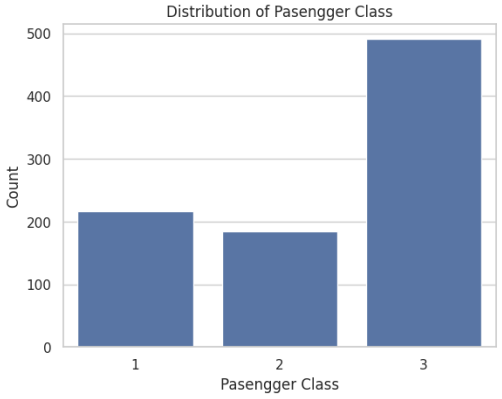
Practice using Google Colab

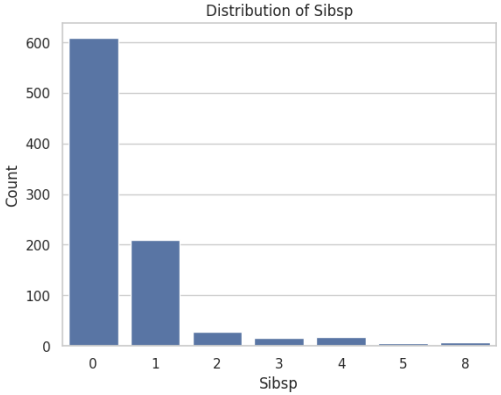
Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

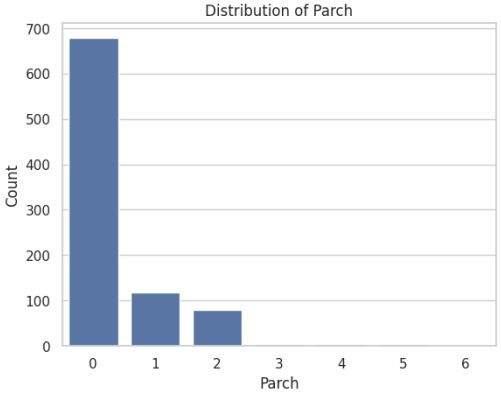
```
[90] # visualizing the distribution of survived, p_class, sibsp and parch
     sns.countplot(x='survived', data=df)
     plt.title('Distribution of Survival')
     plt.xlabel('Survived')
     plt.ylabel('Count')
     plt.show()
     sns.countplot(x='p_class', data=df)
     plt.title('Distribution of Pasengger Class')
     plt.xlabel('Pasengger Class')
     plt.ylabel('Count')
     plt.show()
     sns.countplot(x='sibsp', data=df)
     plt.title('Distribution of Sibsp')
     plt.xlabel('Sibsp')
     plt.ylabel('Count')
     plt.show()
     sns.countplot(x='parch', data=df)
     plt.title('Distribution of Parch')
     plt.xlabel('Parch')
     plt.ylabel('Count')
     plt.show()
     sns.countplot(x='embarked', data=df)
     plt.title('Distribution of Embarked')
     plt.xlabel('Embarked')
     plt.ylabel('Count')
     plt.show()
```

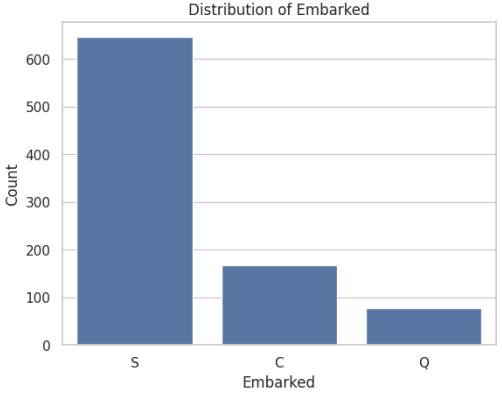
Practice using Google Colab











Owner : Muhammad Aziz Habiburrahim

#RintisKarirImpian

MySkill IdScore

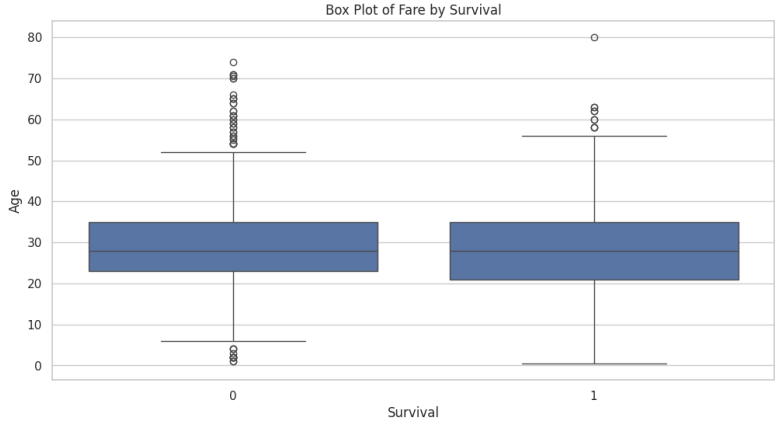
Practice using Google Colab

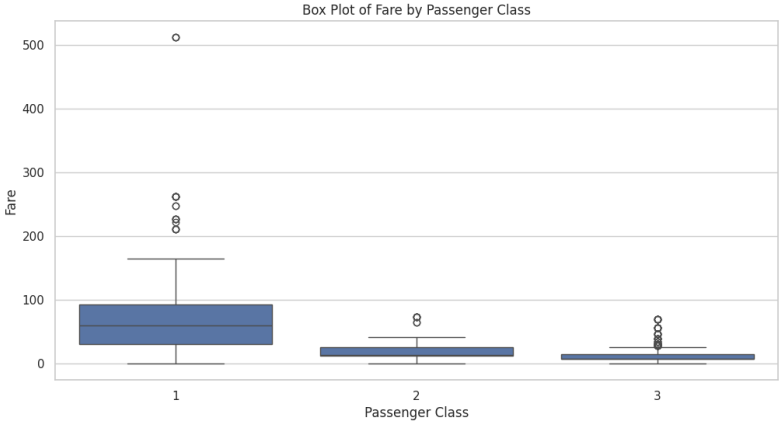
Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
[91] # box plot for numerical features
     plt.figure(figsize=(12, 6))
     sns.boxplot(x='survived', y='age', data=df)
     plt.title('Box Plot of Fare by Survival')
     plt.xlabel('Survival')
     plt.ylabel('Age')
     plt.show()
     plt.figure(figsize=(12, 6))
     sns.boxplot(x='p_class', y='fare', data=df)
     plt.title('Box Plot of Fare by Passenger Class')
     plt.xlabel('Passenger Class')
     plt.ylabel('Fare')
     plt.show()
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.





Owner : Muhammad Aziz Habiburrahim

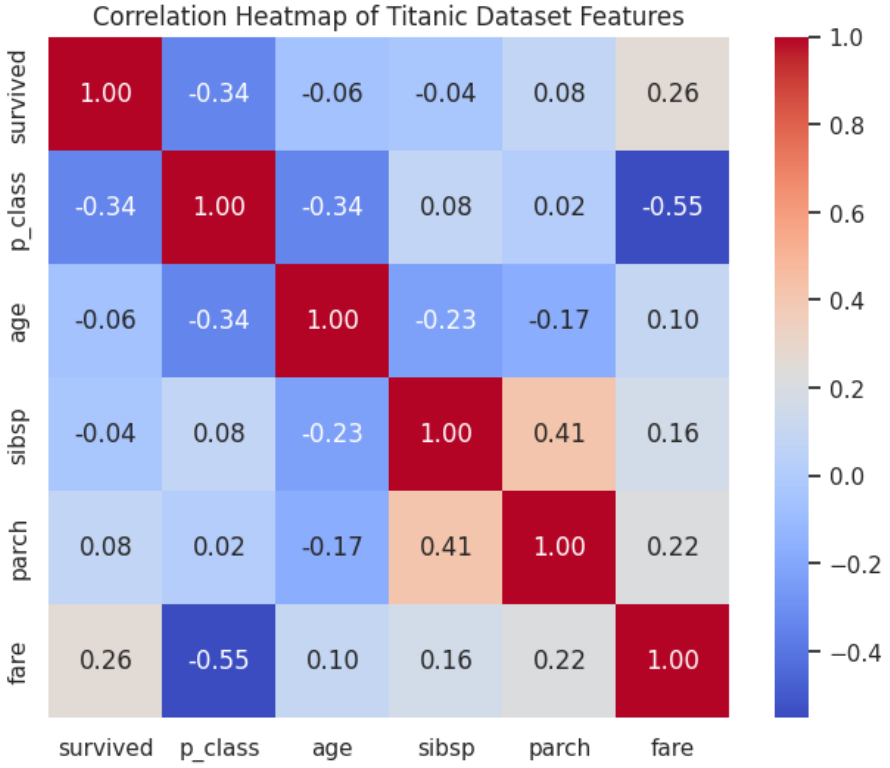
Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.

```
# correlation plot heatmap
# Selecting numerical features for correlation analysis
numerical_features = ['survived', 'p_class', 'age', 'sibsp', 'parch', 'fare']
correlation_matrix = df[numerical_features].corr()
# generating the heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', square=True)
plt.title('Correlation Heatmap of Titanic Dataset Features')
plt.show()
```

Practice using Google Colab

Olah dataset yang sudah disediakan pada link berikut https://bit.ly/DatasetSCDataScienceMar2025MySkillxIdScore pada Google Colab https://colab.research.google.com/ sesuai arahan dari tutor. Jika sudah, screenshot hasil olahan data tersebut dan tambahkan pada slide ini.



Follow me!

Instagram : @mhabibr02

LinkedIn : Muhammad Aziz Habiburrahim

Short Class Data Science and Analysis

Data Analysis Fundamentals