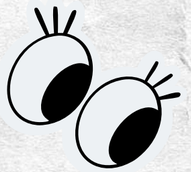
titanic data analysis

Hasbul Wafi

Hello,
I’m Hasbul



Sebagai seorang data enthusiast, saya sedang mencoba menganalisis data korban titanic. Hal ini akan mempermudah strategi evakuasi di masa yang akan mendatang. Dalam proyek ini saya melakukan ***Exploratory Data Analysis (EDA)*** untuk memanipulasi data agar mudah dipahami strukturnya, diidentifikasi masalahnya serta mengvisualisasikan datanya.

Introduction

ibimbing

01. BACKGROUND

02. PROBLEM SOLVING

03. URGENCY

ibimbing

Key Steps

- OBSERVATION DATA
- CLEANSING DATA
- VISUALIZATION DATA

ibimbing

1

Data Observation

Load Data

```
[59]  0s
1  #import data
2  import pandas as pd
3
4  df=pd.read_excel('titanic.xlsx')
5  data = df.copy()
```

✱ Load Data

- Library **pandas** digunakan untuk memanipulasi data
- Membaca file excel bernama **“titanic.xlsx”** dan menyimpannya ke dalam variabel df sebagai **DataFrame.**
- Membuat **salinan utuh** dari DataFrame df ke variabel baru bernama Data.

Data Observation

1

```
# Menampilkan 5 data teratas
data.head()
```

	survived	name	sex	age
0	1	Allen, Miss. Elisabeth Walton	female	29.0000
1	1	Allison, Master. Hudson Trevor	male	0.9167
2	0	Allison, Miss. Helen Loraine	female	2.0000
3	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000
4	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000

✱ Head

data.head() menampilkan 5 baris pertama dari sebuah dataset.

```
# Menampilkan 5 data terbawah
data.tail()
```

	survived	name	sex	age
495	1	Mallet, Mrs. Albert (Antoinette Magnin)	female	24.0
496	0	Mangiavacchi, Mr. Serafino Emilio	male	NaN
497	0	Matthews, Mr. William John	male	30.0
498	0	Maybery, Mr. Frank Hubert	male	40.0
499	0	McCrae, Mr. Arthur Gordon	male	32.0

✱ Tail

data.tail() menampilkan 5 baris terakhir dari sebuah dataset.

ibimbing

Data Observation

1

✱ Random Sample

- Setiap kolom dalam dataset memiliki nilai yang konsisten.
- Kolom survived dan sex hanya berisi nilai biner kategorikal [0,1].
- Kolom name memuat format berupa nama lengkap beserta gelar (seperti Mr., Mrs., Miss.) yang dapat diekstraksi.
- Data tampak bersih tanpa adanya kejanggalan atau masalah.

```
0s
1  # Menampilkan 5 random sampel dari data
2  data.sample(5)
```

	survived	name	sex	age
264	1	Simonius-Blumer, Col. Oberst Alfons	male	56.0
164	1	Homer, Mr. Harry ("Mr E Haven")	male	35.0
109	1	Flynn, Mr. John Irwin ("Irving")	male	36.0
75	0	Colley, Mr. Edward Pomeroy	male	47.0
168	1	Icard, Miss. Amelie	female	38.0

ibimbing

1

Data Observation

✱ Info Data

- Dataset terdiri dari total 500 baris dan 4 kolom.
- Pada kolom age, terdapat 49 nilai yang hilang (diperoleh dari 500 – 451).
- Jenis data sudah sesuai dengan masing-masing kolom.
- Perlu dilakukan pemeriksaan terhadap nilai yang duplikat serta mengatasi missing value pada kolom age.

```
0s
1  #Menampilkan info dari data
2  data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 4 columns):
 #  Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0  survived 500 non-null   int64
 1  name     500 non-null   object
 2  sex      500 non-null   object
 3  age      451 non-null   float64
dtypes: float64(1), int64(1), object(2)
memory usage: 15.8+ KB
```

ibimbing

1

Data Observation

```
[66]  1  #Melihat kolom apa saja yang ada pada excel
      2  data.columns
```

```
Index(['survived', 'name', 'sex', 'age'], dtype='object')
```

```
[69]  1  #Mengelompokkan data jenis katergori dan numerik
      2  categoricals=['name','sex']
      3  numericals=['survived','age']
```

✱ Data Columns

- Dataset ini memiliki 4 kolom: 'survived', 'name', 'sex', dan 'age'.
- Kolom-kolom tersebut dapat dikelompokkan berdasarkan jenisnya:
  1.Kategorikal: 'name', 'sex'
  2.Numerik: 'survived', 'age'.

ibimbing

1

Data Observation

```
0s
1  #Menampilkan statistical summary dari data numerik
2  data[numericals].describe()
```

	survived	age
count	500.000000	451.000000
mean	0.540000	35.917775
std	0.498897	14.766454
min	0.000000	0.666700
25%	0.000000	24.000000
50%	1.000000	35.000000
75%	1.000000	47.000000
max	1.000000	80.000000

✱ Statistical Summary

- Kolom survived memiliki data biner, di mana 0 menunjukkan penumpang yang tidak selamat dan 1 menunjukkan yang selamat.
- Sekitar 54% dari total penumpang berhasil selamat (mean = 0.54).
- Pada kolom age, terdapat beberapa data yang hilang, dengan hanya 451 dari total 500 entri yang tersedia.
- Usia penumpang berkisar dari bayi termuda (0.66 tahun) hingga individu tertua (80 tahun).
- Distribusi usia terlihat cukup seimbang, karena nilai mean dan median hampir sama, yaitu sekitar 35 tahun.

ibimbing

1

Data Observation

✱ Categorical Summary

- Kolom name sebagian besar bersifat unik, dengan 499 dari 500 entri berbeda, meskipun terdapat kemungkinan nilai duplikat karena frekuensi name sebanyak 2.
- Kolom sex hanya memiliki dua kategori, yaitu **laki-laki (male)** dan **perempuan (female).**
- Laki-laki mendominasi jumlah penumpang, yaitu sebanyak 288 dari total 500 entri.
- Kolom-kolom kategorikal tidak memiliki nilai yang hilang (missing values).

```
0s
1  #Menampilkan categorical summary dari data kategori
2  data[categoricals].describe()
```

	name	sex
count	500	500
unique	499	2
top	Eustis, Miss. Elizabeth Mussey	male
freq	2	288

ibimbing

1

Data Observation

```
0s
#Menampilkan detail dari data numerik
for col in numericals:
    print(data[col].value_counts())
```

```
survived
1  270
0  230
Name: count, dtype: int64
age
24.0000  23
30.0000  20
36.0000  19
18.0000  14
42.0000  14
```

✱ Statistical Details

- Data survived memiliki sifat biner dengan distribusi yang hampir merata, yaitu 270 dan 230.
- Kolom age mencakup 73 nilai unik dari total 500 entri, menunjukkan keberagaman usia penumpang.
- Usia yang paling sering muncul adalah 24, 30, dan 36 tahun.
- Sebagian besar nilai age hanya muncul satu kali, sehingga dapat memengaruhi skewness (kemiringan) saat data divisualisasikan.

ibimbing

1

Data Observation

```
#Menampilkan detail dari data kategori
for col in categoricals:
  print(data[col].value_counts())
```

```
name
Eustis, Miss. Elizabeth Mussey                2
Becker, Miss. Ruth Elizabeth                  1
Becker, Miss. Marion Louise                   1
Becker, Master. Richard F                     1
Beauchamp, Mr. Henry James                    1
Beane, Mrs. Edward (Ethel Clarke)             1
Beane, Mr. Edward                             1
Bateman, Rev. Robert James                    1
Banfield, Mr. Frederick James                 1
Ball, Mrs. (Ada E Hall)                       1
```

✱ Categorical Details

- Kemungkinan adanya duplikasi data terlihat dari nama "Eustis, Miss. Elizabeth Mussey," yang muncul sebanyak dua kali.
- Pada kolom sex, distribusi tidak seimbang karena jumlah penumpang laki-laki jauh lebih banyak dibandingkan dengan perempuan.

ibimbing

2 Data Cleansing

```
[74]  1  len(data)
```

```
500
```

```
1  len(data.drop_duplicates())
```

```
499
```

```
[80]  1  duplicates=data[data.duplicated(keep=False)]
      2  duplicates
```

	survived	name	sex	age
104	1	Eustis, Miss. Elizabeth Mussey	female	54.0
349	1	Eustis, Miss. Elizabeth Mussey	female	54.0

✱ Duplicated Data

- Dataset memiliki satu data duplikat, terlihat dari jumlah data unik yang hanya 499 dari total 500.
- Duplikasi terjadi pada nama "Eustis, Miss. Elizabeth Mussey," yang muncul dua kali dengan informasi identik pada indeks 104 dan 349.
- Penting untuk menghapus data duplikat tersebut agar hasil analisis tetap akurat dan bebas dari bias.

ibimbing

Data Cleansing

2

```
dup_count = duplicates.groupby(list(data.columns)).size().reset_index(name = "duplicates count")

sorted_duplicates=dup_count.sort_values(by = ['duplicates count'], ascending = False)

sorted_duplicates
```

	survived	name	sex	age	duplicates count
0	1	Eustis, Miss. Elizabeth Mussey	female	54.0	2

```
#Menghilangkan data yang duplikat
data = data.drop_duplicates()
```

```
len(data)
```

```
499
```

✱ Handling Duplicates

- Frekuensi kemunculan data duplikat dihitung dan diurutkan berdasarkan jumlahnya.
- Semua data yang teridentifikasi sebagai duplikat telah dihapus.
- Setelah penghapusan, jumlah total data berkurang dari 500 menjadi 499.

ibimbing

Data Cleansing

2

Null Values

```
1  #Melihat Jumlah Value null untuk setiap kolom
2  data.isna().sum()
```

0s

	0
survived	0
name	0
sex	0
age	49

dtype: int64

✱ Null Values

- Metode **data.isna().sum()** digunakan untuk menghitung jumlah nilai kosong atau null pada dataset.
- Ditemukan sebanyak 49 nilai null pada kolom age, sesuai dengan informasi yang telah diketahui sebelumnya.
- Sementara itu, semua kolom lainnya terisi sepenuhnya tanpa nilai yang hilang.

ibimbing

Data Cleansing

2

Fill Null Values

```
[87]  1  print(data['age'].dtype)
      2  print(data['age'].median())
```

```
float64
35.0
```

```
[89]  1  for column in data.select_dtypes(include=['number']).columns:
      2      data[column] = data[column].fillna(data[column].median())
```

```
<ipython-input-89-3fa0d82d15ea>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
    data[column] = data[column].fillna(data[column].median())
```

✱ Fill Null Values

- Karena tipe data pada kolom age bersifat numerik, nilai yang kosong diisi menggunakan **median** dari kolom tersebut, yaitu 35.
- Setelah proses pengisian selesai, kolom age tidak lagi mengandung nilai kosong atau null.

ibimbing

Data Cleansing

2

```
0s
1  #Menampilkan info data setelah proses handling duplicates
2  data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 499 entries, 0 to 499
Data columns (total 4 columns):
 #  Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0  survived  499 non-null  int64
 1  name  499 non-null  object
 2  sex  499 non-null  object
 3  age  499 non-null  float64
dtypes: float64(1), int64(1), object(2)
memory usage: 19.5+ KB
```

✱ Info Data (Final)

Setelah proses penghapusan data duplikat selesai, dataset telah terbebas dari nilai yang terduplikasi. Selain itu, data yang kosong pada kolom age telah diisi dengan nilai median, yaitu 35, karena kolom ini bersifat numerik dan menggunakan median dianggap sebagai pendekatan yang tepat untuk menjaga keselarasan distribusi data.

Visualization Data

3


```
[33]  1  import matplotlib.pyplot as plt
      2  import seaborn as sns
      3  sns.set_theme(style="whitegrid")
      4  plt.figure(figsize=(8, 5))
      5  ax = data['sex'].value_counts().plot( kind='bar', color=sns.
         color_palette("pastel"),edgecolor='black')
      6  plt.title("Jumlah Penumpang Berdasarkan Jenis Kelamin",
         fontsize=14, fontweight='bold')
      7  plt.xlabel("Jenis Kelamin", fontsize=12)
      8  plt.ylabel("Jumlah Penumpang", fontsize=12)
      9  for bar in ax.patches:ax.annotate(f'{bar.get_height()}',(bar.get_x
         () + bar.get_width() / 2, bar.get_height()),ha='center',
         va='bottom', fontsize=10, color='black')
     10  plt.tight_layout()
     11  plt.show()
```

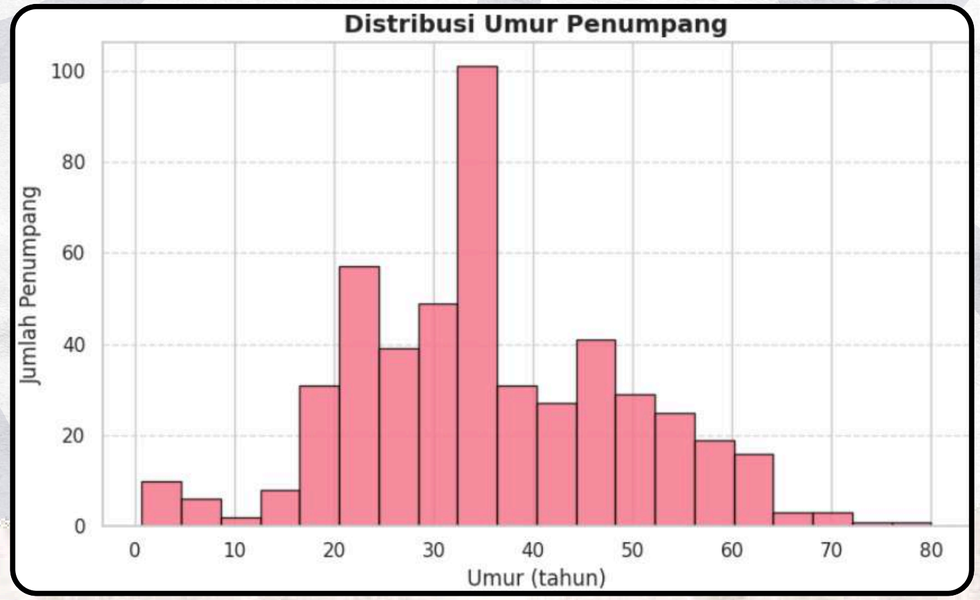
✱ Amount & Gender

Dari grafik, terlihat bahwa jumlah penumpang laki-laki lebih banyak dibandingkan dengan jumlah penumpang perempuan:

- Penumpang laki-laki: Sebanyak 288 orang.
- Penumpang perempuan: Sebanyak 211 orang.

Visualization Data

3



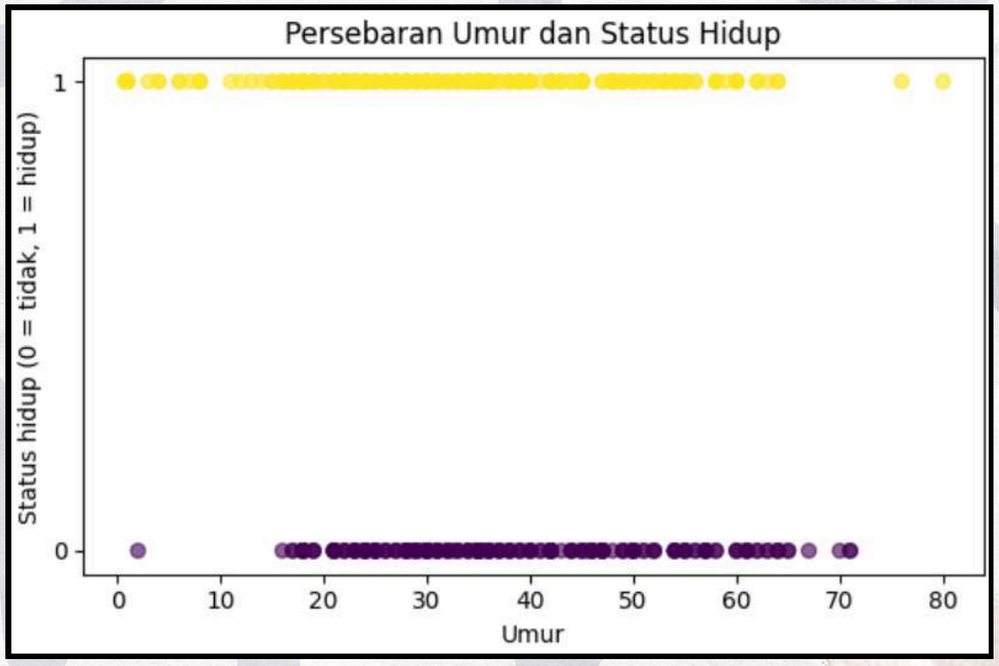

```python
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_theme(style="whitegrid")
plt.figure(figsize=(8, 5))
plt.hist(data['age'], bins=20, color=sns.color_palette("husl", 1)[0], edgecolor='black', alpha=0.8)
plt.title("Distribusi Umur Penumpang", fontsize=14, fontweight='bold')
plt.xlabel("Umur (tahun)", fontsize=12)
plt.ylabel("Jumlah Penumpang", fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

✱ Age Distribution

Berdasarkan histogram distribusi umur diatas, dapat dilihat bahwa mayoritas penumpang berada dalam rentang usia 30 hingga 40 tahun, dengan puncak jumlah penumpang di sekitar usia 35 tahun.

ibimbing

3




```
1  plt.figure(figsize=(6,4))
2  plt.scatter(data['age'], data['survived'], alpha= 0.6, c=data
   ['survived'])
3  plt.title("Persebaran Umur dan Status Hidup")
4  plt.xlabel("Umur")
5  plt.ylabel("Status hidup (0 = tidak, 1 = hidup)")
6  plt.yticks([0,1])
7  plt.tight_layout()
8  plt.show()
```

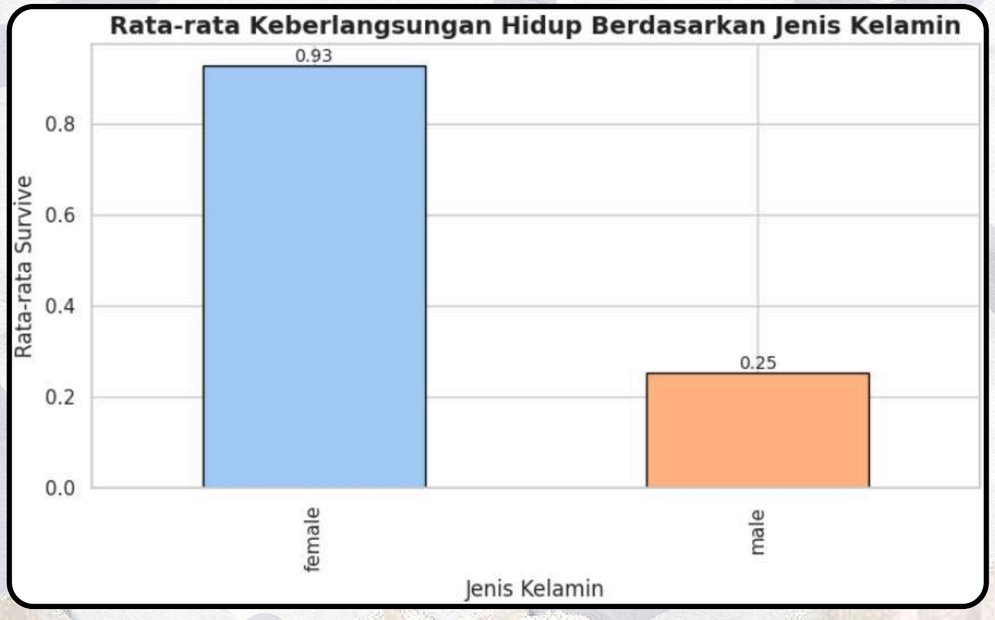
Visualization Data

✱ Age & Living Status

Dilihat dari scatterplot tersebut, persebaran data untuk status tidak hidup banyak berada di sekitar umur 20-70 (outlier untuk 0-10),
sedangkan untuk hidup itu tersebar banyak di umur 0-60 (outlier untuk 70-80).

Visualization Data

3




```
[31]
1  import matplotlib.pyplot as plt
2  import seaborn as sns
3
4  sns.set_theme(style="whitegrid")
5  plt.figure(figsize=(8, 5))
6  ax = data.groupby('sex')['survived'].mean().plot(
7      kind='bar',
8      color=sns.color_palette("pastel"),
9      edgecolor='black'
10 )
11 plt.title("Rata-rata Keberlangsungan Hidup Berdasarkan Jenis
   Kelamin", fontsize=14, fontweight='bold')
12 plt.xlabel("Jenis Kelamin", fontsize=12)
13 plt.ylabel("Rata-rata Survive", fontsize=12)
14 for bar in ax.patches:
15     ax.annotate(f'{bar.get_height():.2f}',
16                 (bar.get_x() + bar.get_width() / 2, bar.get_height
                    ()),
17                 ha='center', va='bottom', fontsize=10)
18 plt.tight_layout()
19 plt.show()
```

✱ Average Survival

Berdasarkan grafik, bahwa rata-rata keberlangsungan hidup atau "survive" untuk perempuan ("female") jauh lebih tinggi dibandingkan dengan laki-laki ("male").

- Perempuan memiliki rata-rata survive sebesar 0.93.
- Laki-laki hanya memiliki rata-rata survive sebesar 0.25.

Kesimpulan

Berdasarkan projek ini, dapat disimpulkan beberapa hal mengenai data ini, diantaranya:

AGE

Umur memengaruhi tingkat keselamatan penumpang Titanic karena anak-anak lebih banyak diselamatkan dibandingkan pria dewasa, sementara lansia menghadapi tantangan mobilitas yang lebih besar selama evakuasi.

GENDER

Faktor gender memainkan peran signifikan dalam tingkat keselamatan korban Titanic, di mana perempuan memiliki peluang lebih besar untuk selamat dibandingkan laki-laki, yang kemungkinan dipengaruhi oleh penerapan aturan "women and children first" selama proses evakuasi.

SURVIVED

Perempuan memiliki tingkat keberlangsungan hidup ("survive") yang jauh lebih tinggi dibandingkan dengan laki-laki pada situasi yang dimaksud. Rata-rata survive perempuan mencapai 0.93, sedangkan rata-rata survive laki-laki hanya 0.25.

Thank You

Let’s Work Together!

 @has.fie_

 sebul1306

 hasbulwafi1306@gmail.com