TITANIC DATA ANALYSIS

Created by Livia Junike to fulfill the assignment of Digital Skill Fair 38.0

DATA OBSERVATION

Load Data

```
1  import pandas as pd
2
3  df = pd.read_excel('titanic.xlsx')
4  data = df.copy()
   ✓ [1] 900ms
```

- 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.**

Head & Tail

```
1 data.head()
[7]
```

5 rows 5 rows × 4 cols

	survived	name	sex	age
0	1	Allen, Miss. Elisabeth…	female	29.0000
1	1	Allison, Master. Hudso…	male	0.9167
2	0	Allison, Miss. Helen L…	female	2.0000
3	0	Allison, Mr. Hudson Jo…	male	30.0000
4	0	Allison, Mrs. Hudson J…	female	25.0000

data.head() menampilkan 5 baris pertama dari dataset

```
1 data.tail()
[9]
```

5 rows 5 rows × 4 cols

	survived	name	sex	age
495	1	Mallet, Mrs. Albert (A…	female	24.0
496	0	Mangiavacchi, Mr. Sera…	male	NaN
497	0	Matthews, Mr. William …	male	30.0
498	0	Maybery, Mr. Frank Hub…	male	40.0
499	0	McCrae, Mr. Arthur Gor…	male	32.0

data.tail() menampilkan 5 baris terakhir dari dataset

Random Sample

```
data.sample(5) #mengambil random sample dari data
[14]
```

5 rows 5 rows × 4 cols

	survived	name	sex	age
190	1	Longley, Miss. Gretche…	female	21.0
34	0	Borebank, Mr. John Jam…	male	42.0
247	1	Rothschild, Mrs. Marti…	female	54.0
68	1	Chevre, Mr. Paul Romai…	male	45.0
433	1	Harris, Mr. George	male	62.0

- Semua kolom memiliki isi yang sesuai.
- Kolom survived dan sex hanya berisi dua nilai (kategorikal biner [0,1])
- Pada head dan tail terdapat beberapa nilai ekstrem atau outlier.
- Kolom nama mengandung format : nama lengkap dan gelar (Mr., Mrs., Miss.), yang dapat diekstrak
- Data terlihat baik dan tidak ada anomali.

Info Data

```
1 data.info()
[16]
```

```
<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
```

Observasi :

- Dataset berisi 500 baris dan 4 kolom.
- Kolom age memiliki 49 missing values (didapat dari 500 – 451).
- Tipe data sudah sesuai dengan isi kolom.
- Perlu cek nilai yang double dan penanganan missing value pada age.

Data Columns

```
data.columns #melihat kolom apa saja yang ada pada excel
[19]
```

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

```
#kelompokan untuk data jenis kategori dan numerik
categoricals = ['name', 'sex']
numericals = ['survived', 'age']
[21]
```

- Data terdiri dari 4 kolom: 'survived', 'name', 'sex', dan 'age'.
- Data kemudian dikelompokkan berdasarkan tipe:
  - **Kategorikal** : 'name', 'sex'
  - **Numerik** : 'survived', 'age'

Statistical Summary

```
#perintah ini untuk menunjukkan statistical summary dari data numerik
data[numericals].describe()
```

[23]

8 rows 8 rows × 2 cols

	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

Observasi :

- Kolom survived hanya berisi nilai biner (0 = tidak selamat, 1 = selamat).
- Sekitar 54% penumpang selamat (mean = 0.54).
- Kolom age memiliki missing values (hanya 451 dari 500 data).
- Usia penumpang bervariasi dari bayi (0.66) hingga lanjut usia (80 tahun).
- Distribusi usia cukup merata, dilihat dari mean dan median yang hampir sama yakni (~35)

Categorical Summary

```
1 data[categoricals].describe()
[25]
```

4 rows 4 rows × 2 cols

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

Observasi :

- Kolom name hampir seluruhnya unique (499 dari 500), kemungkinan ada nilai duplikat, dilihat dari frekuensi name ada 2.
- Kolom sex hanya punya 2 kategori: male dan female.
- Mayoritas penumpang adalah laki-laki (288 dari 500).
- Tidak ada missing values pada kolom kategorikal.

Statistical Details

```
for col in numericals:
    print(data[col].value_counts())
```

[31]

```
survived
1    270
0    230
Name: count, dtype: int64
age
24.0    23
30.0    20
36.0    19
18.0    14
45.0    14
        ..
11.0     1
59.0     1
80.0     1
2.0      1
3.0      1
Name: count, Length: 73, dtype: int64
```

Observasi :

- survived adalah data biner dengan distribusi cukup seimbang (270 & 230).
- Kolom age memiliki 73 nilai unik dari 500 data yang menunjukkan age ini beragam, dengan usia paling umum adalah 24, 30, dan 36 tahun.
- Banyak value age yang muncul hanya satu kali, ini memengaruhi kemiringan jika data divisualisasikan.

Categorical Details

```
for col in categoricals:
    print(data[col].value_counts())
```

[29]

```
name
Eustis, Miss. Elizabeth Mussey                         2
Allen, Miss. Elisabeth Walton                          1
Angle, Mrs. William A (Florence "Mary" Agnes Hughes)   1
Becker, Miss. Ruth Elizabeth                           1
Becker, Miss. Marion Louise                            1
                                                      ..
Holverson, Mr. Alexander Oskar                         1
Hogeboom, Mrs. John C (Anna Andrews)                   1
Hippach, Mrs. Louis Albert (Ida Sophia Fischer)        1
Hippach, Miss. Jean Gertrude                           1
McCrae, Mr. Arthur Gordon                              1
Name: count, Length: 499, dtype: int64
sex
male      288
female    212
Name: count, dtype: int64
```

Observasi :

- Ada kemungkinan data duplikat, karena ada nama “Eustis, Miss. Elizabeth Mussey” yang berjumlah 2.
- Distribusi tidak merata pada kolom sex, karena persebaran male lebih besar daripada female.

DATA CLEANSING

Duplicated Data

```
1 len(data)
[32]
```

500

```
1 len(data.drop_duplicates())
[34]
```

499

```
1 duplicates = data[data.duplicated(keep=False)]
2 duplicates
[38]
```

2 rows 2 rows × 4 cols

	survived	name	sex	age
104	1	Eustis, Miss. Elizabet…	female	54.0
349	1	Eustis, Miss. Elizabet…	female	54.0

Observasi :

- Terdapat 1 data duplikat karena data uniknya hanya 499 dari 500
- Duplikat berasal dari data ”Eustis, Miss. Elizabeth Mussey” yang muncul dua kali dengan data yang sama (index 104 & 349).
- Data duplikat ini perlu dihapus agar analisis tidak bias.

Handling Duplicates

```
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
```

[49]

1 row · 1 rows × 5 cols

	survived	name	sex	age	duplicates count
0	1	Eustis, Miss. Elizabet…	female	54.0	2

```
data = data.drop_duplicates()

len(data)
```

[52]

499

- Hitung kemunculan frekuensi duplikat dan diurutkan.
- Menghapus seluruh data duplikat.
- Panjang data berubah dari 500 menjadi 499 setelah data duplikat dihapus.

Null Values

```
#melihat jumlah value null untuk setiap kolom
data.isna().sum() #bisa juga gunakan data.isnull().sum()
```

[60]

```
survived     0
name         0
sex          0
age         49
dtype: int64
```

Info data

```
0   survived   500 non-null    int64
1   name       500 non-null    object
2   sex        500 non-null    object
3   age        451 non-null    float64
```

- Menggunakan data.isna.sum() untuk menghitung nilai yang kosong atau null
- Ternyata ada 49 data null pada kolom age, sesuai dengan info data yang tersedia. Sedangkan kolom yang lain terisi semua.

Fill Null Values

```
print(data['age'].dtype)
print(data['age'].median())
[65]
```

```
float64
35.0
```

```
for column in data.select_dtypes(include=['number']).columns:
    data[column] = data[column].fillna(data[column].median())
[80]
```

```
data.isna().sum()
[81]
```

```
survived    0
name        0
sex         0
age         0
dtype: int64
```

- Karena data type kolom age itu numerical, maka kita isi data yang kosong menggunakan nilai median dari kolom age tersebut, yakni bernilai 35.
- Setelah pengisian nilai, maka kolom age sudah tidak ada nilai yang kosong atau null.

Info Data (Final)

```
1  data.info()
   [82]
```

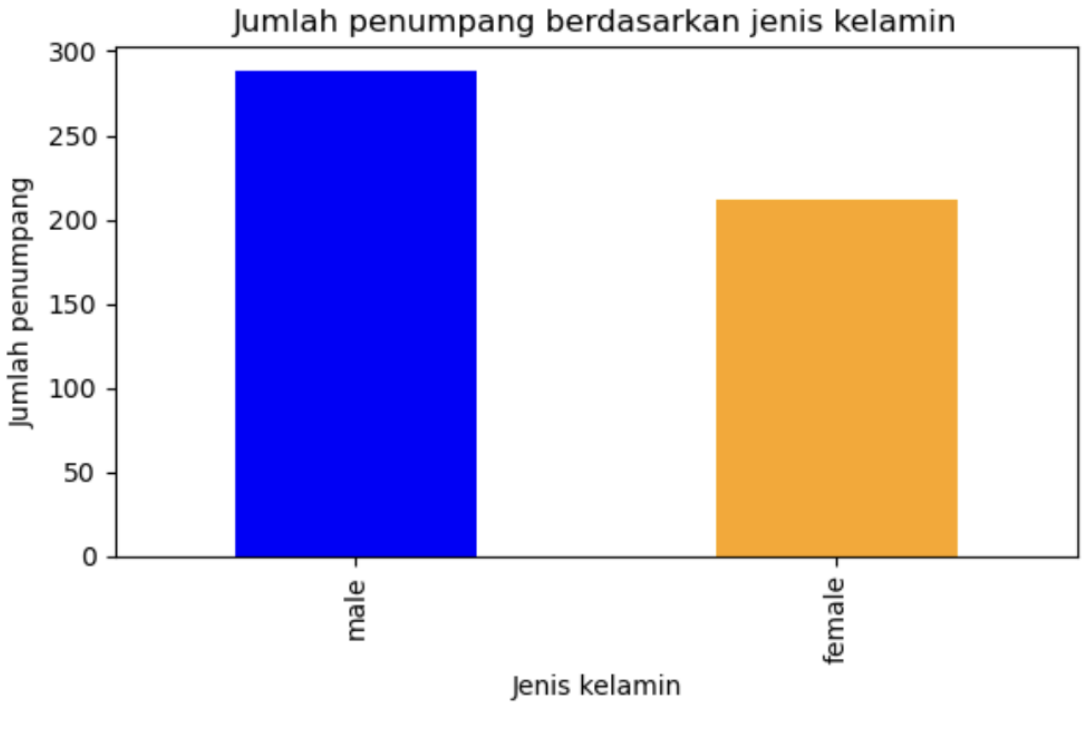
```
<class 'pandas.core.frame.DataFrame'>
Index: 499 entries, 0 to 499
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   survived  499 non-null    int64
 1   name      499 non-null    object
 2   sex       499 non-null    object
 3   age  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

Observasi :

- Setelah pembersihan data duplikat dan pengisian data yang kosong, semua kolom bersih dan lengkap (semuanya sudah sesuai).

DATA VISUALIZATION

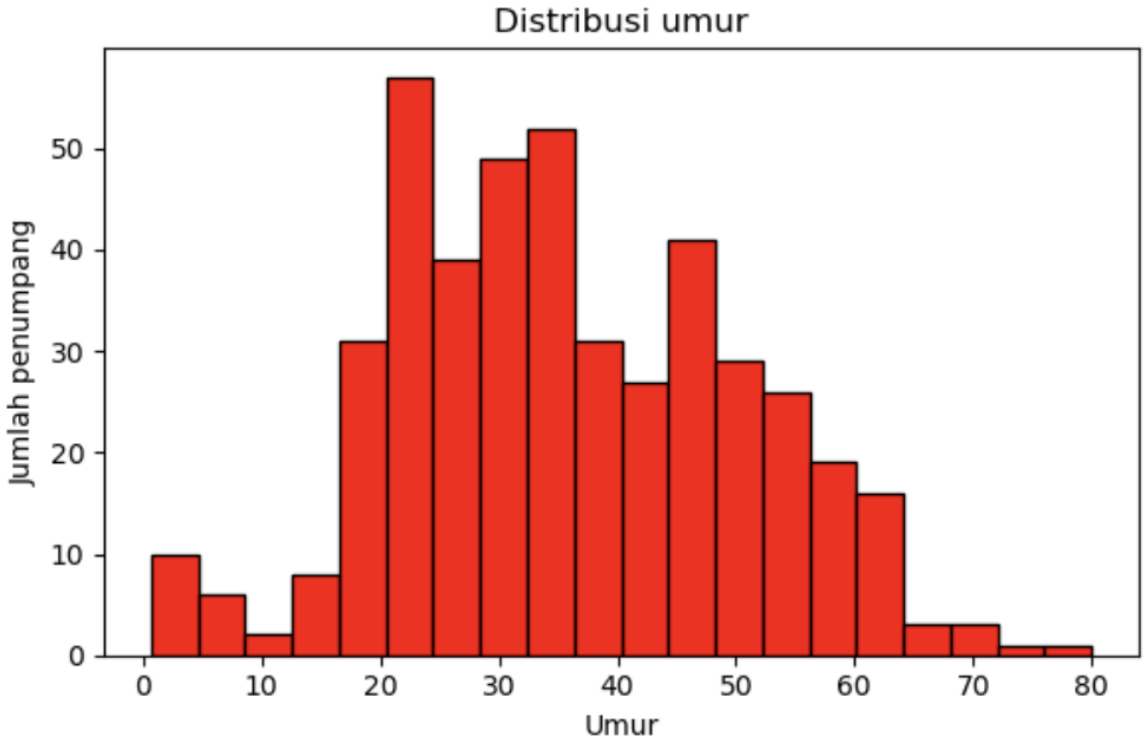
Amount & Gender



Observasi :

- Dilihat dari visualisasi, jenis kelamin (sex) untuk male lebih tinggi dibandingkan female (hampir 100 data).

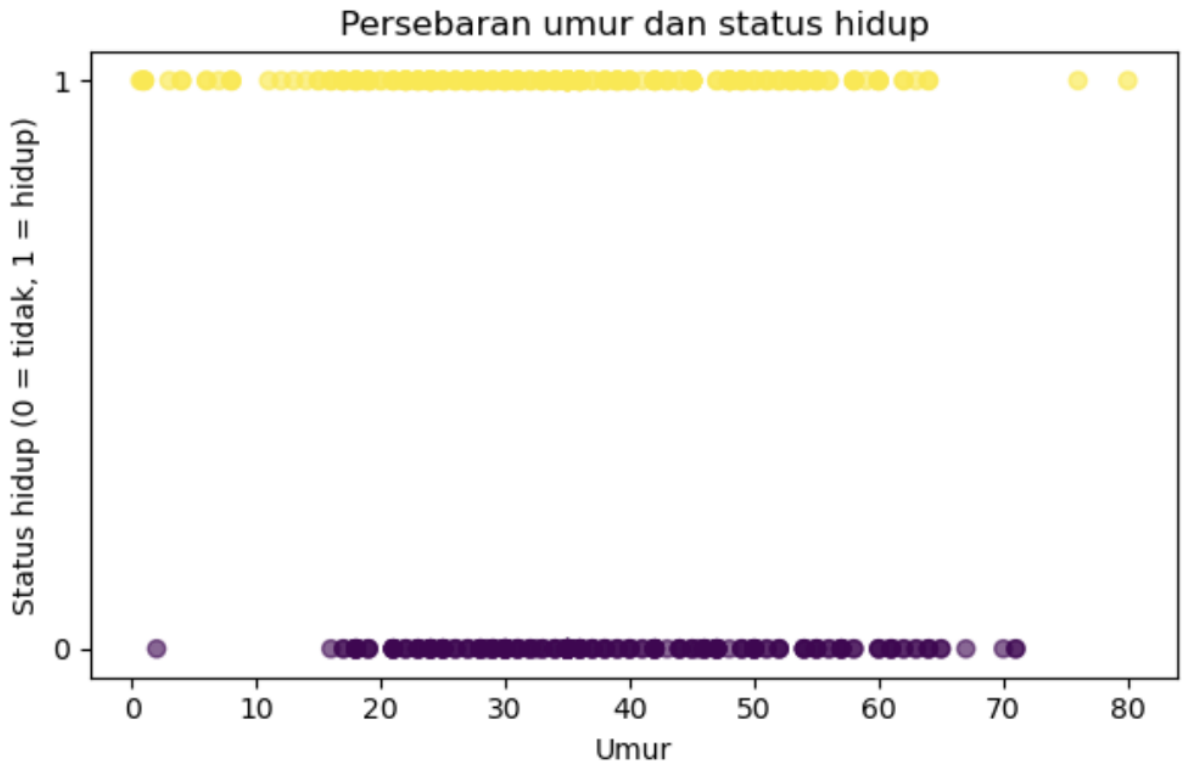
Age Distribution



Observasi :

- Dilihat dari histogram, penumpang titanic tersebut paling banyak berusia 20 hingga 35 tahun. Ada beberapa outlier untuk usia 70 keatas.

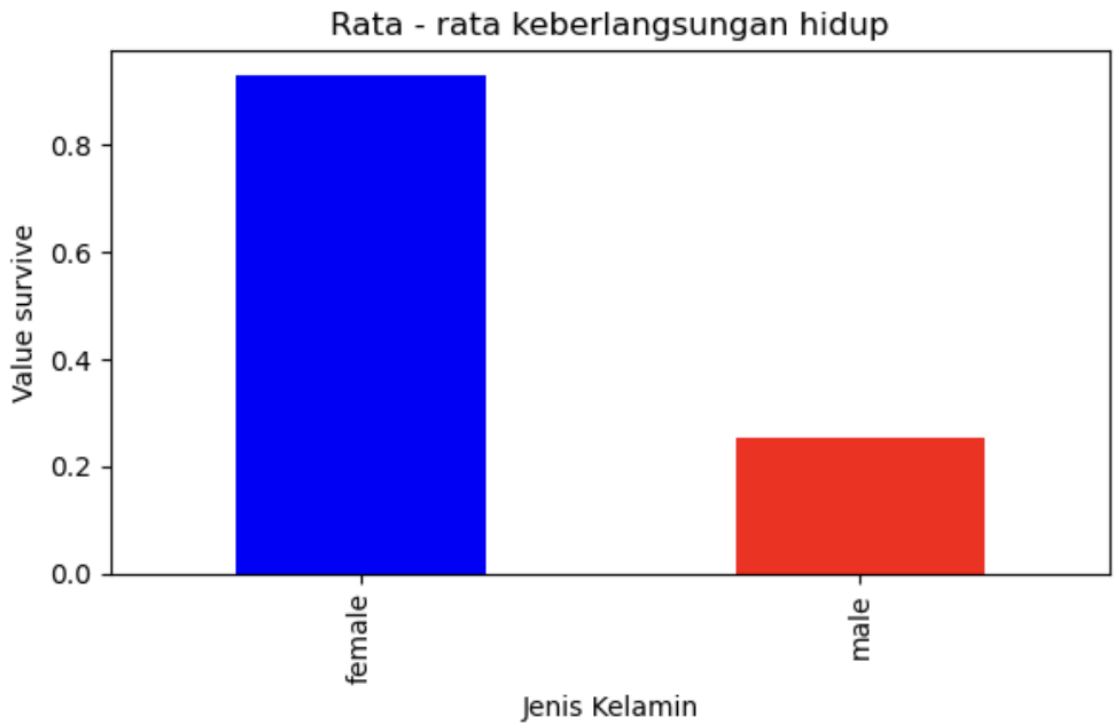
Age & Living Status



Observasi :

- 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).

Average Survival



Observasi :

- Dilihat dari visualisasi, jenis kelamin female memiliki rata – rata keberlangsungan hidup yang jauh lebih tinggi dibandingkan male (bahkan female hampir mencapai 100%, sedangkan male hanya sekitar 20% saja)

Livia Junike