

Understanding your Data

Exploratory Data Analysis using Python

Press Space for next page →



Reza Rizky



[rezarzky](#)



[rezarzky](#)

Bakhtiar A.

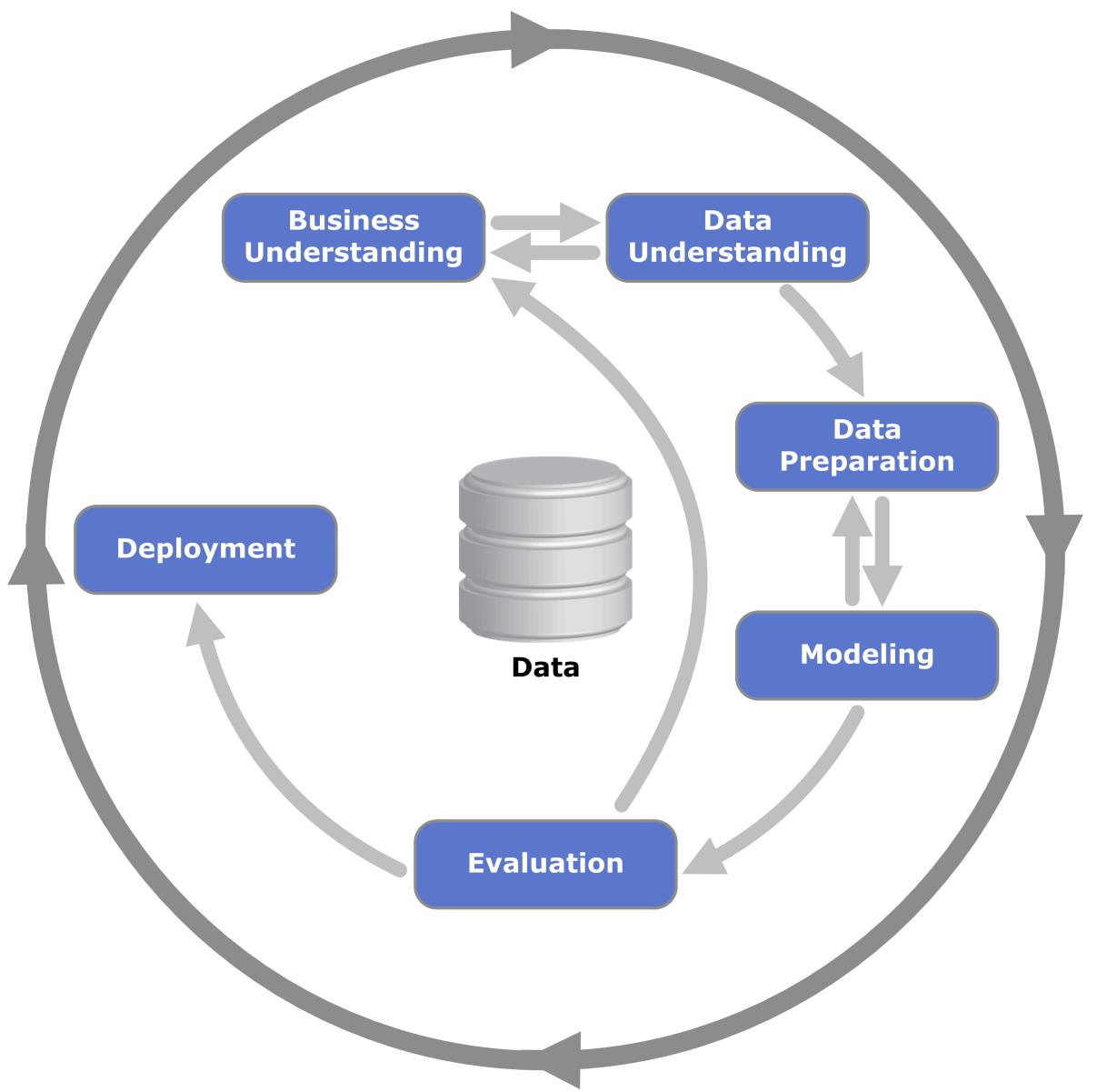


[maziyank](#)



[maziyank](#)

EDA ?



Exploratory Data Analusis (EDA) adalah seperangkat pendekatan statistik untuk mengeksplorasi dan memahami data dengan menggunakan teknik-teknis statistik dan visualisasi.



Manfaat EDA:

- Memahami permasalahan yang ada pada data sehingga dapat merencanakan langkah-langkah data cleansing.
- Merumuskan pertanyaan-pertanyaan bisnis lain, sehingga dapat mempertajam pertanyaan bisnis yang merupakan hasil dari step data understanding
- Menghasilkan output berupa hasil analisis deskriptif yang dapat disajikan pada visualisasi data dan dashboard.



3 Jenis EDA:

- **Univariate Analysis**
- **Bivariate Analysis**
- **Multivariate Analysis**

Melakukan EDA menggunakan python sangatlah mudah karena python memiliki segudang libraries untuk analisis statistik dan visualisasi (e.g numpy, pandas, matplotlib, etc).

Univariate Analysis

- Analisis univariat mengeksplorasi variabel (atribut) satu per satu. Variabel dapat berupa **kategorikal** atau **numerik**.
 - Variabel numerik dapat diubah menjadi bagian kategorikal dengan proses yang disebut **binning** atau **diskritisasi**.
 - Dimungkinkan juga untuk mengubah variabel kategorikal menjadi pasangan numeriknya dengan proses yang disebut **encoding**.

```
139     title="Instagram"
140     target="_blank"
141     rel="noopener noreferrer">
142     href={trackUrl('Instagram')}
143   >
144   Instagram
145   </a>
146 </ul>
147 </ul>
148 </div>
149 );
150 }
151
152 renderWhatShowLinks() {
153   return (
154     <div className={style.links}>
155       <ul className={style.list}>
156         <li className={style.item}>
157           {this.renderFooterMain()}
158           {this.renderFooterSub()}
159           {this.renderFooterSlogan()}
160           {this.renderFooterSlogan()}
161           {this.renderFooterSlogan()}
162           {this.renderFooterSlogan()}
163           {this.renderFooterSlogan()}
164           {this.renderFooterSlogan()}
165         </ul>
166       </div>
167     );
168   }
169
170 renderWhatShowItem(title, url) {
171   return (
172     <li className={styles.footer}>
173       <a
174         href={trackUrl(url)}
175         target="_blank"
176         rel="noopener noreferrer">
177         >
178         {title}
179       </a>
180     </li>
181   );
182 }
183
184 renderFooterSub() {
185   return (
186     <div className={styles.footerSub}>
187       <a href="/" title="Home - Unsplash">
188         <Icon
189           type="logo"
190           className={styles.footerSubLogo}
191         />
192       </a>
193       <span className={styles.footerSlogan}>
194     </div>
195   );
196 }
197
198 render() {
199   return (
200     <footer className={styles.footerGlobal}>
201       <div className="container">
202         {this.renderFooterMain()}
203         {this.renderFooterSub()}
204       </div>
205     </footer>
206   );
207 }
208 }
209
```

Resep Univariate

- Cek informasi data (dimensi data, tipe data, jumlah data)
- Cek missing value
- Analisis Data Categorical
 - Count, %Count
 - Pie Chart, Bar Chart
- Analisis Data Numerical
 - Min, Max, Mean, Median, Mode
 - Range, Quantiles, Variance, Standard Deviation, Coefficient of Variation
 - Skeweness, Kurtosis
 - Histogram, Box Plot



Cek Informasi Data

Menggunakan cara yang telah dipelajari pada materi pengantar Pandas.

```
# mengetahui dimensi DataFrame  
df.shape  
  
# jumlah data (baris x kolom) pada dataframe  
df.size  
  
# list kolom pada dataframe  
df.columns  
  
# list index pada dataframe  
df.index  
  
# informasi dataframe, kolom tipe data dsb.  
df.info()  
  
#informasi tipe data pada dataframe  
df.dtypes
```



Tipe Data pada Dataset Titanic

PassengerId	int64
Survived	int64
Pclass	int64
Name	object
Sex	object
Age	float64
SibSp	int64
Parch	int64
Ticket	object
Fare	float64
Cabin	object
Embarked	object
dtype: object	

Setelah kita mengetahui tipe data masing-masing kolom, selanjutnya kita bisa menentukan pendekatan analisis datanya.

Tipe Data Pandas

Pandas dtype	Python type	Usage
object	str or mixed	Text or mixed numeric and non-numeric values
int64	int	Integer numbers
float64	float	Floating point numbers
bool	bool	True/False values
datetime64	datetime	Date and time values
timedelta[ns]	NA	Differences between two datetimes
category	NA	Finite list of text values

Kenapa Tipe Data Penting?

- Dalam satu kolom terdapat beberapa tipe data (mixed).
 - Secara visual kita melihat data numerik, ternyata setelah di cek tipenya object.
 - Tipe Data menentukan langkah analisis berikutnya.

Studi Kasus Sales Data

Cek Missing Value

```
# Memeriksa missing value  
df.isna().sum()  
  
# Memeriksa missing value %  
data.isna().sum() * 100 / len(data)
```

Contoh:

```
PassengerId      0  
Survived         0  
Pclass           0  
Name             0  
Sex              0  
Age            177  
SibSp            0  
Parch            0  
Ticket           0  
Fare             0  
Cabin          687  
Embarked        2  
dtype: int64
```

Cek Missing Value Secara Visual

```
# Instal library missing no  
!pip install missingno
```

```
import missingno as msno  
  
# dalam bentuk matrix  
msno.matrix(df)  
  
# dalam bentuk heatmap  
msno.heatmap(df)  
  
# dalam bentuk bar plot  
msno.bar(df)
```

Analisis Data Kategorikal

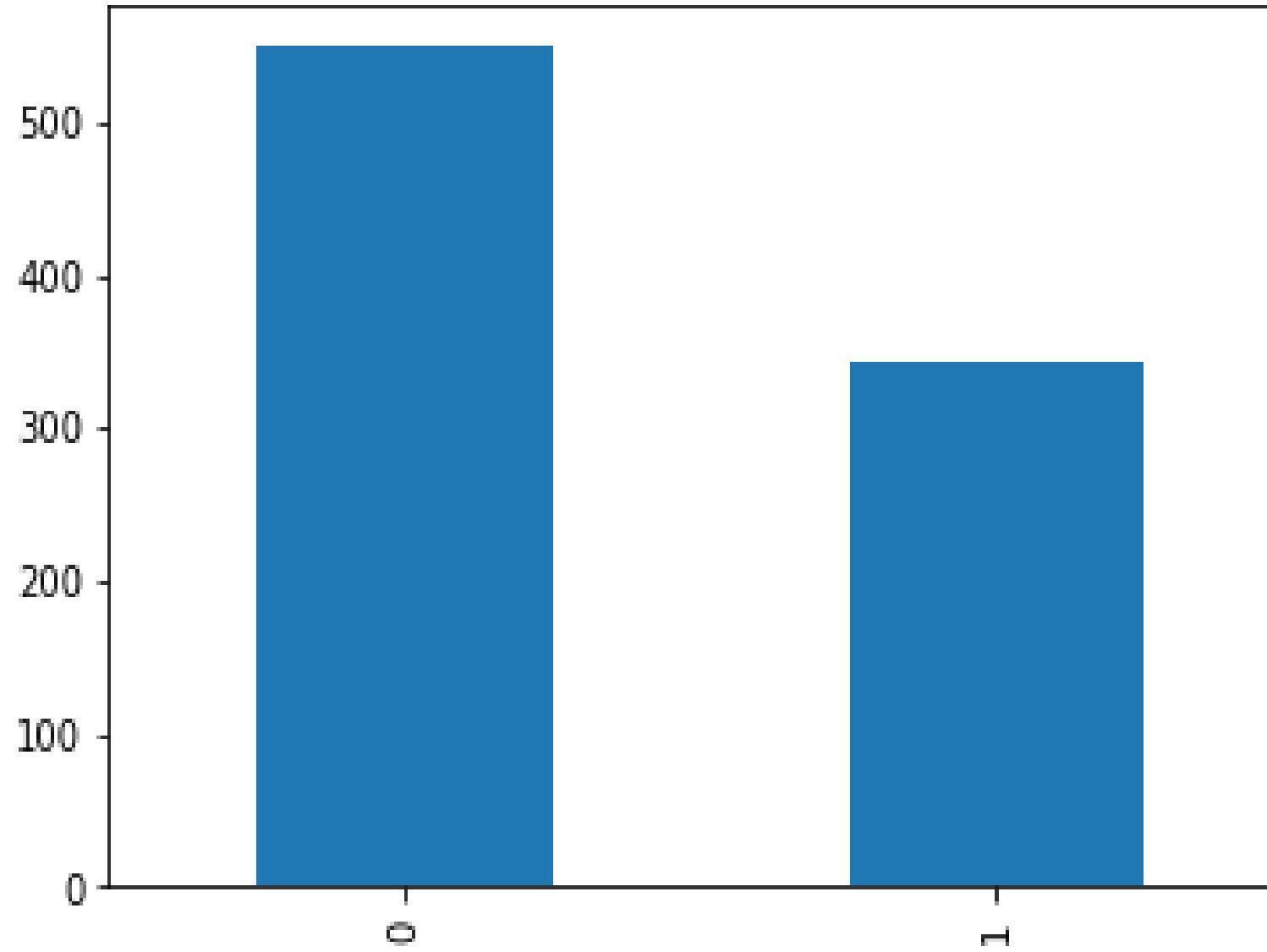
Stats	Visualization	Description
Count	Bar Chart	The number of values of the specified variable.
%Count	Pie Chart	The percentage of values of the specified variable.

Implementasi:

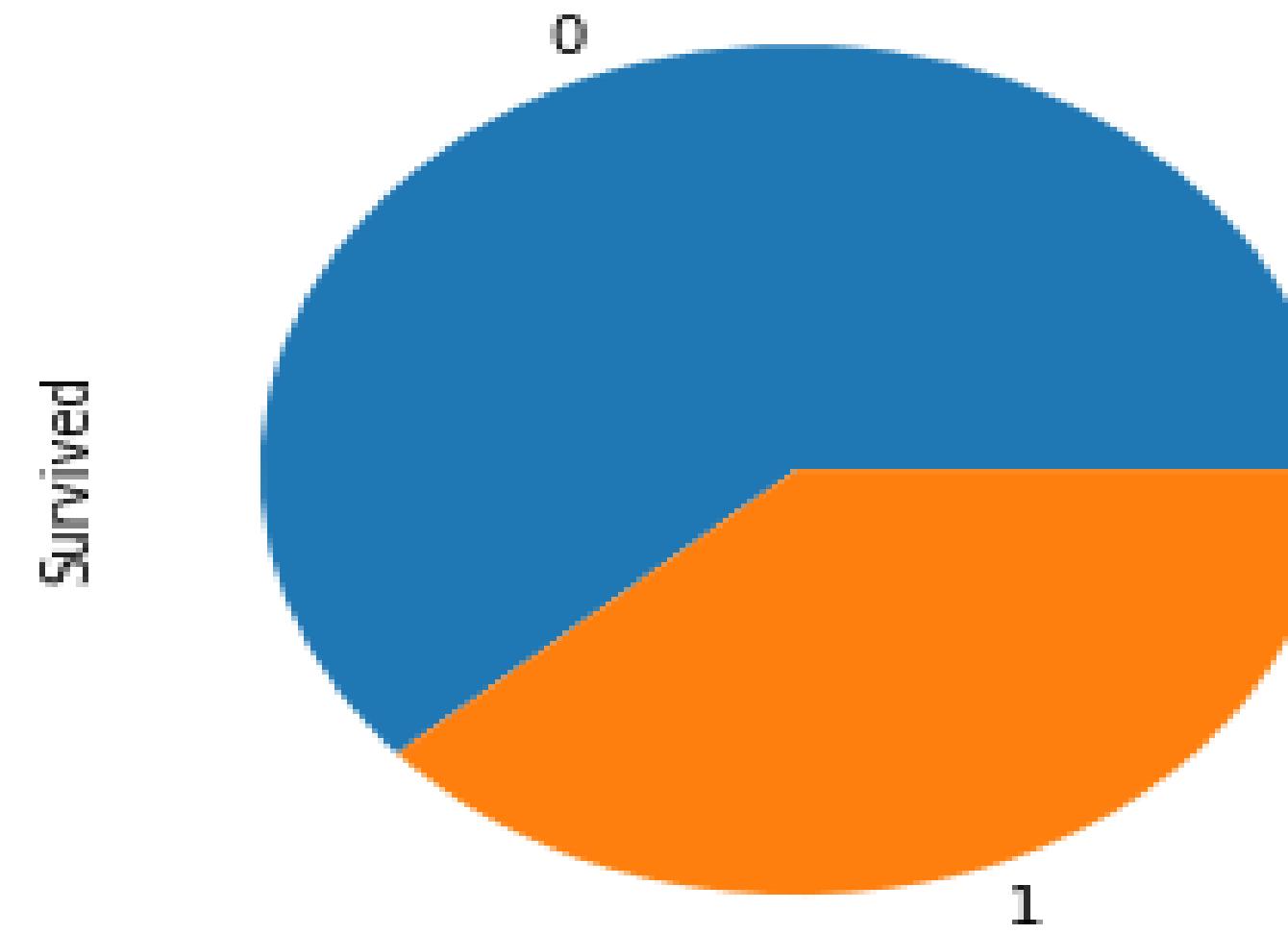
```
# Menghitung Jumlah Nilai per Kolom  
df.Survived.value_counts()  
  
# Menghitung Persentase Nilai per Kolom  
df.Survived.value_counts(normalize=True)
```

Visualisasi Data Kategorikal

```
# Membandingkan Nilai per Kolom  
# menggunakan Visualisasi Bar Chart  
df.Survived.value_counts().plot.bar()
```



```
# Membandingkan Nilai per Kolom  
# menggunakan Visualisasi Pie Chart  
df.Survived.value_counts(normalize=True).plot.pie()
```



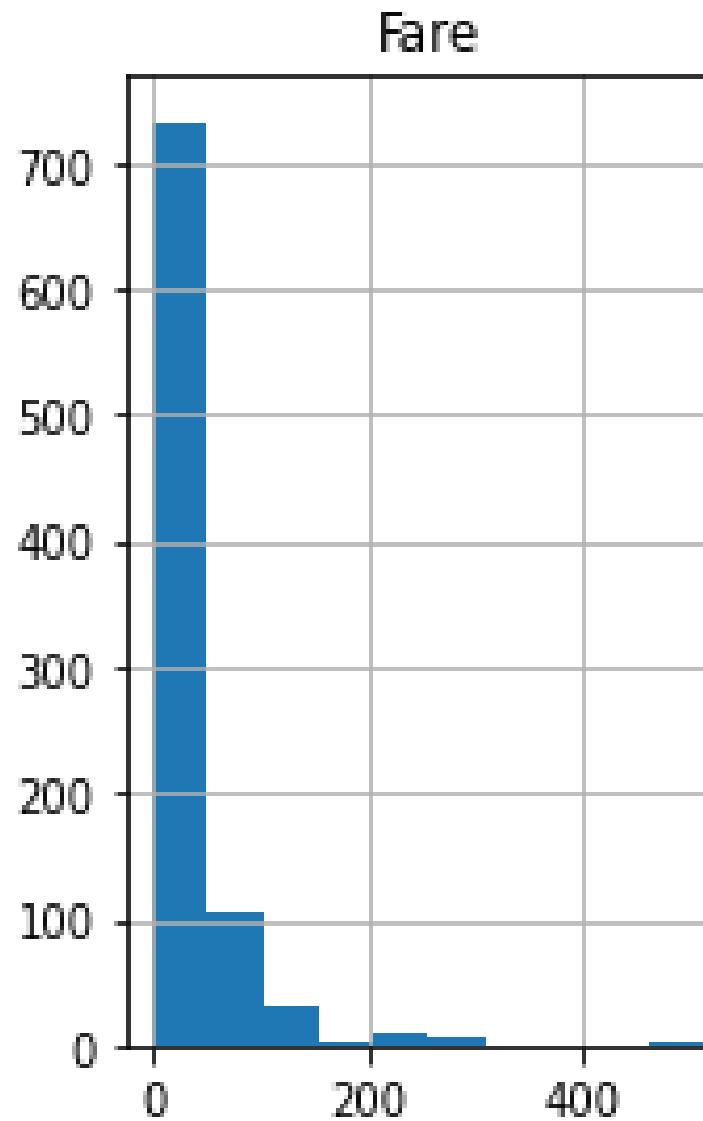
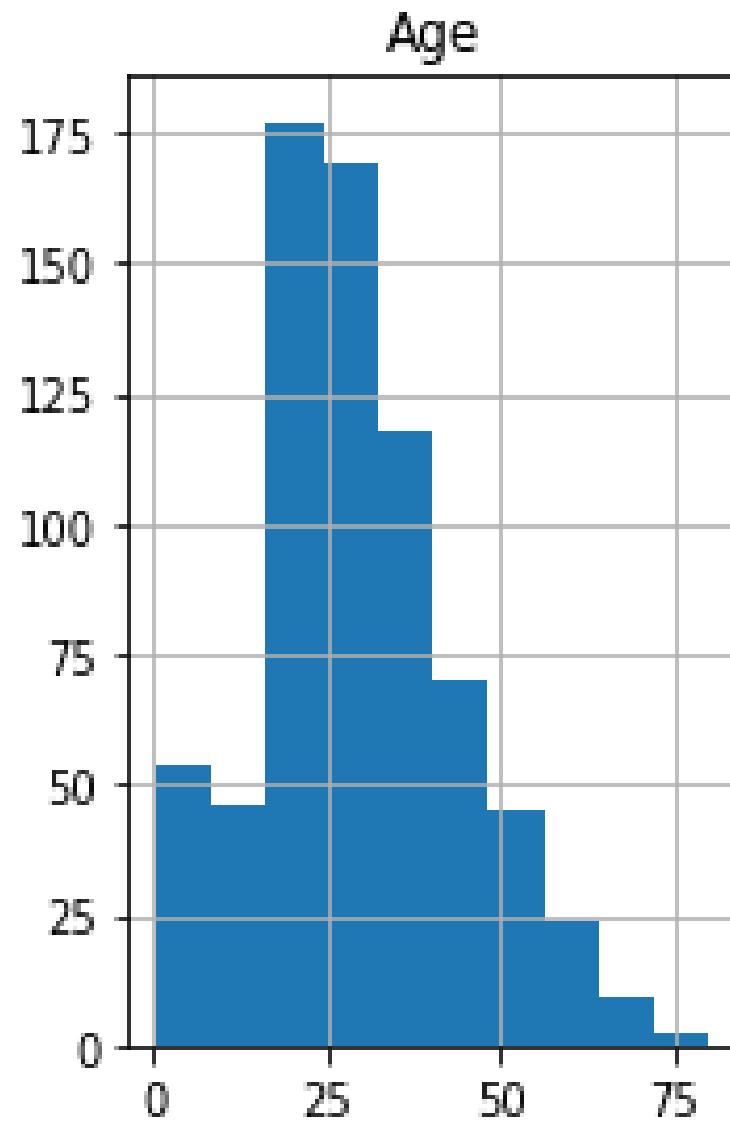
Analisis Data Numerical

- Min, Max, Mean, Median, Mode
- Range, Quantiles, Variance, Standard Deviation, Coefficient of Variation
- Skeweness, Kurtosis
- Visualisasi : Histogram, Box Plot

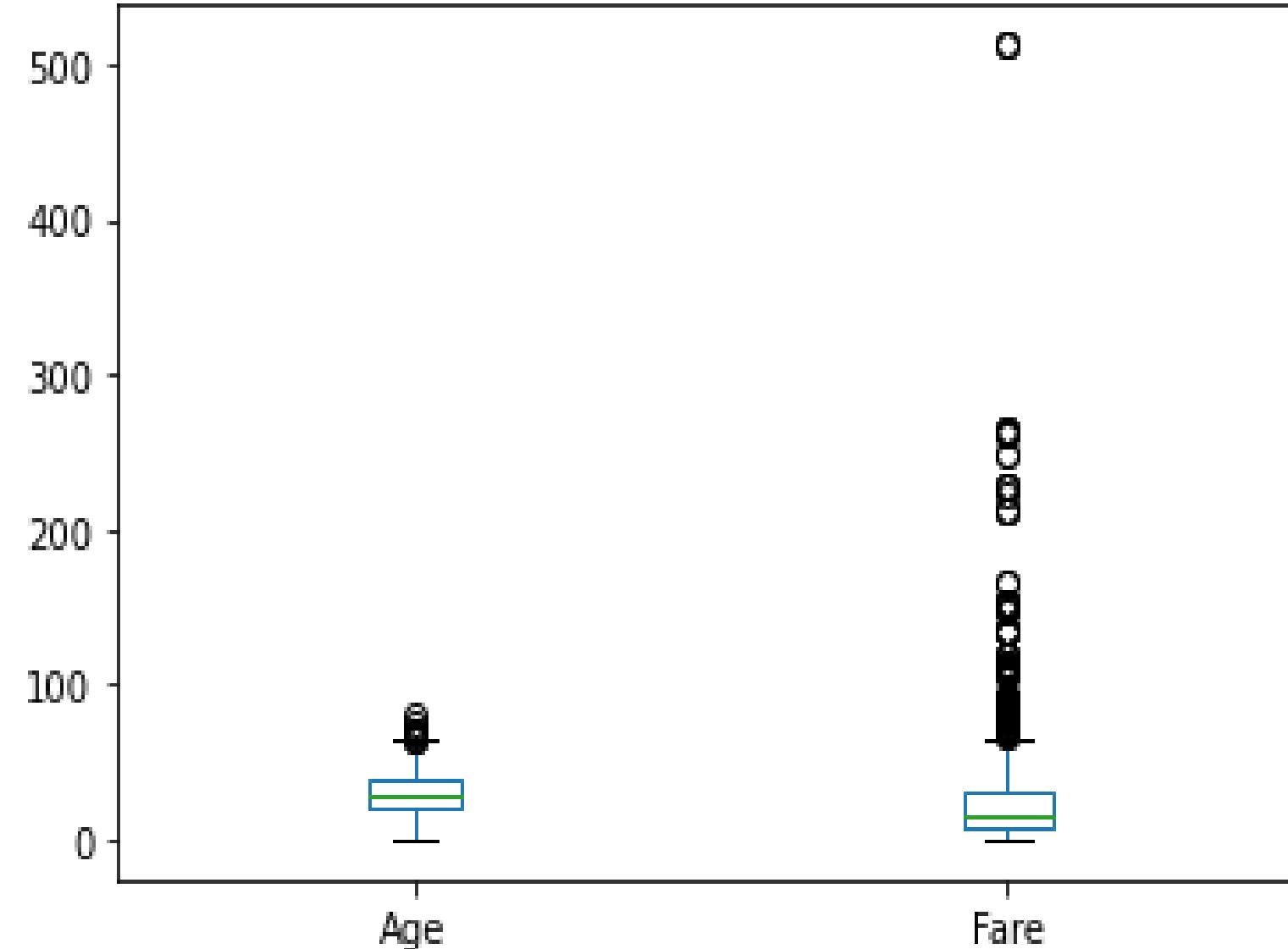
```
# Untuk (Min, Max, Mean, Median, Quantiles, Standar Deviasi) kita hanya perlu menggunakan fungsi "describe()".  
df.describe()  
  
# Variance  
df[['Age', 'Fare']].var()  
  
# Skewness  
df[['Age', 'Fare']].skew()  
  
#Kurtosis  
df[['Age', 'Fare']].kurtosis()  
  
#Range  
df.Age.max() - df.Age.min()
```

Visualisasi Data Numerikal

```
#histogram plot  
df[['Age', 'Fare']].hist()
```

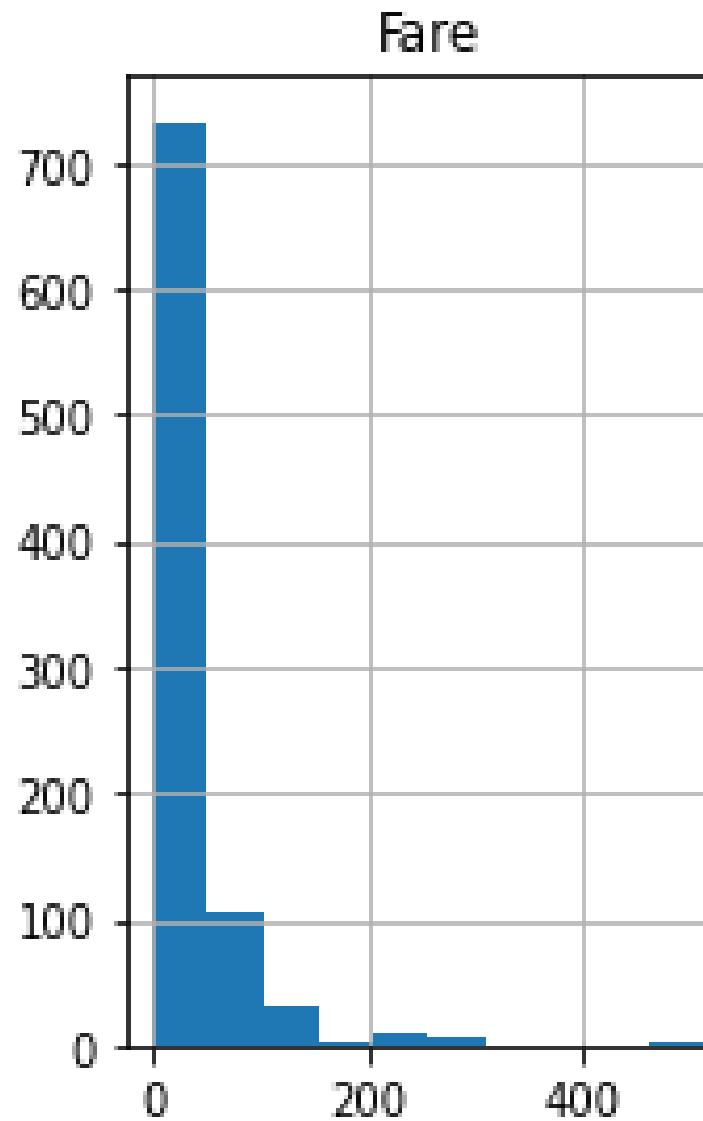
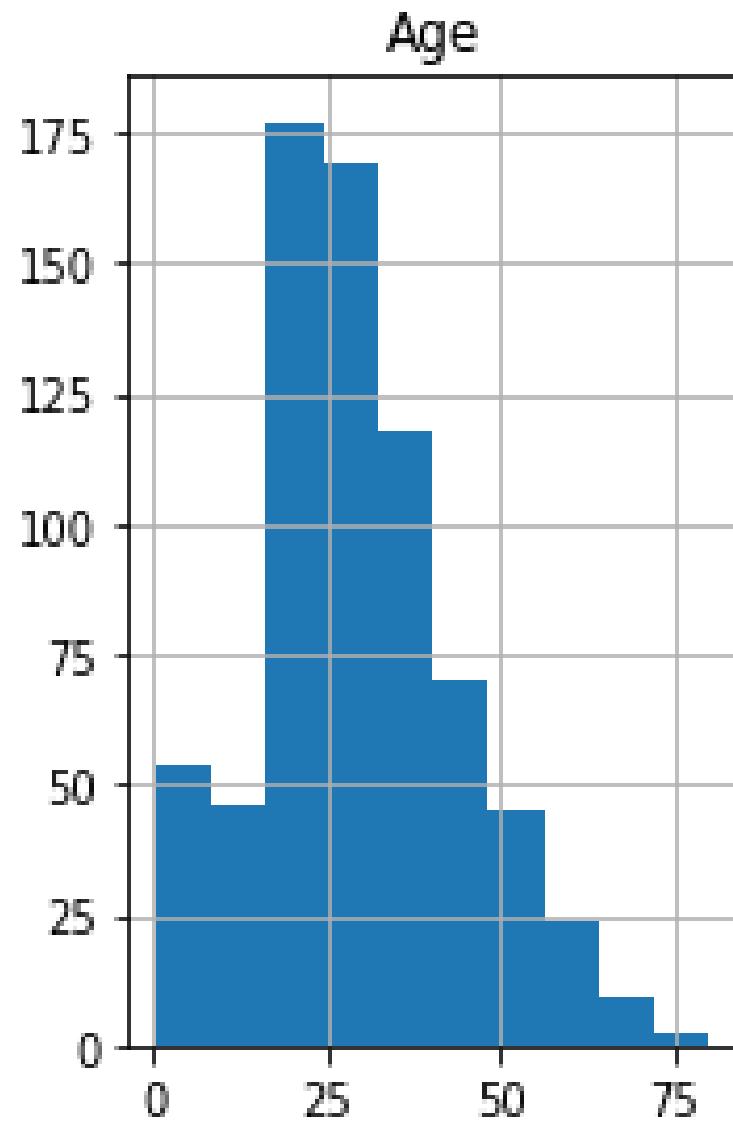


```
#Box Plot  
df[['Age', 'Fare']].plot.box()
```

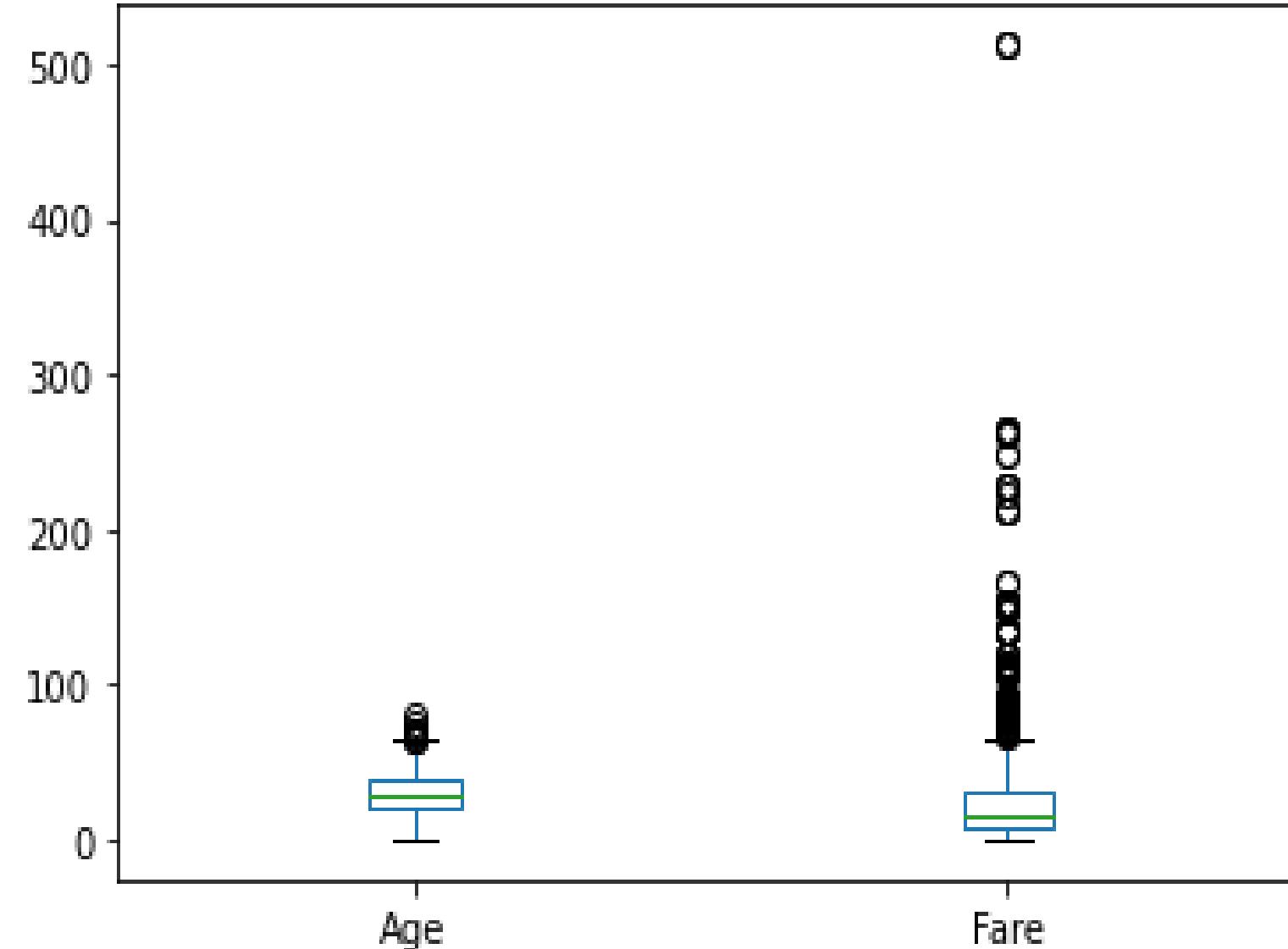


Visualisasi Data Numerikal

```
#histogram plot  
df[['Age', 'Fare']].hist()
```



```
#Box Plot  
df[['Age', 'Fare']].plot.box()
```



Thank You!

Today is enough.