

Topic: Customizing Visualizations

- Added annotations, adjusted figure sizes, and used advanced legends.
- Example: Annotated key points in a scatter plot.

A scatter plot uses dots to represent values for two different numeric variables. In Python, we have a library matplotlib in which there is a function called scatter that helps us to create Scatter Plots. Here, we will use matplotlib.pyplot.scatter() method to plot.

Syntax : `matplotlib.pyplot.scatter(x,y)`

Parameters:

- *x and y are float values and are the necessary parameters to create a scatter plot*
- *marker : MarkerStyle, default: rcParams[“scatter.marker”] (default: ‘o’)*
- *cmap : cmapstr or Colormap, default: rcParams[“image.cmap”] (default: ‘viridis’)*
- *linewidths : float or array-like, default: rcParams[“lines.linewidth”] (default: 1.5)*
- *alpha : float, default: None → represents the transparency*

Annotation of matplotlib means that we want to place a piece of text next to the scatter. There can be two cases depending on the number of the points we have to annotate :

1. Single point annotation
2. All points annotation

Single Point annotation

In single-point annotation we can use matplotlib.pyplot.text and mention the x coordinate of the scatter point and y coordinate + some factor so that text can be distinctly visible from the plot, and then we have to mention the text.

Syntax: `matplotlib.pyplot.text(x, y, s)`

Parameters:

- *x, y : scalars — The position to place the text. By default, this is in data coordinates. The coordinate system can be changed using the transform parameter.*
- *s : str — The text.*
- *fontsize — It is an optional parameter used to set the size of the font to be displayed.*

Approach:

1. Import libraries.
2. Create data.
3. Make scatter plot.
4. Apply `plt.text()` method.

25-11-2024

Training Day – 47

Topic: Final Data Analysis

- Conducted descriptive and inferential analyses on the final dataset.
- Example: Analyzed correlations between variables using `.corr()`.

After cleaning and combining datasets, the next critical step in the data analysis process is to conduct both **descriptive** and **inferential analyses** to uncover meaningful insights and relationships within the data. This step helps summarize the data and make predictions or inferences based on it. Below, we'll cover key techniques used in final data analysis.

1. Descriptive Analysis

Descriptive statistics summarize and describe the characteristics of the dataset. This includes measures of central tendency (mean, median, mode), dispersion (variance, standard deviation), and the distribution of variables.

- **Key Metrics:**
 - **Mean:** The average of a dataset.
 - **Median:** The middle value when data is sorted.
 - **Mode:** The most frequently occurring value.
 - **Standard Deviation:** Measures the spread of data points around the mean.
 - **Variance:** The square of the standard deviation.
 - **Skewness:** Measures the asymmetry of the distribution.
 - **Kurtosis:** Measures the "tailedness" of the distribution.
- **Example: Descriptive Statistics in Python**

```
import pandas as pd

# Sample dataset
data = pd.DataFrame({
    'Age': [23, 45, 22, 34, 40],
    'Salary': [45000, 54000, 47000, 58000, 60000]
})

# Descriptive statistics
descriptive_stats = data.describe()
print(descriptive_stats)
```

Output:

```
shell
Copy code
      Age      Salary
count  5.000000   5.000000
mean   32.800000  52800.000000
std     8.460517  5907.926474
min    22.000000  45000.000000
25%    23.000000  47000.000000
50%     34.000000  54000.000000
75%     40.000000  58000.000000
max     45.000000  60000.000000
```

2. Analyzing Correlations Between Variables

Understanding the relationships between variables is crucial in data analysis. **Correlation** is a statistical measure that expresses the extent to which two variables are linearly related. The correlation coefficient ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 meaning no correlation.

- **Pearson Correlation:** Measures the linear relationship between two continuous variables.
- **Spearman Rank Correlation:** Used for ordinal data or when the relationship between variables is not linear.
- **Example: Correlation Analysis**

```
import pandas as pd

# Sample dataset
data = pd.DataFrame({
    'Age': [23, 45, 22, 34, 40],
    'Salary': [45000, 54000, 47000, 58000, 60000],
    'Experience': [1, 10, 2, 8, 12]})
# Calculate correlation matrix
corr_matrix = data.corr()
print(corr_matrix)
```

Output:

Markdown

	Age	Salary	Experience
Age	1.000000	0.967858	0.822845
Salary	0.967858	1.000000	0.970010
Experience	0.822845	0.970010	1.000000

From the output, we can see:

- The **Salary** and **Experience** variables are highly positively correlated with each other (0.97).
- There is a strong positive correlation between **Age** and **Salary** (0.97), indicating that older individuals in this sample tend to have higher salaries.

3. Inferential Analysis

Inferential analysis involves making predictions or inferences about a population based on a sample. This typically involves hypothesis testing, regression analysis, and confidence intervals. Key techniques include:

- **Hypothesis Testing:**
 - **Null Hypothesis (H0):** A statement of no effect or no difference.
 - **Alternative Hypothesis (H1):** The statement that there is an effect or difference.

- **P-value:** Used to assess the strength of the evidence against the null hypothesis (usually, $p < 0.05$ is considered statistically significant).
 - **t-tests / ANOVA:** Used to compare means between groups.
- **Regression Analysis:**
 - **Linear Regression:** Used to predict the value of a dependent variable based on one or more independent variables.
 - **Logistic Regression:** Used when the dependent variable is categorical (e.g., binary classification).

Topic: Creating a Dashboard

- Integrated multiple Matplotlib visualizations into one figure.
- Example: Combined a line chart, bar chart, and pie chart in subplots.

Matplotlib allows you to combine multiple visualizations (such as line charts, bar charts, and pie charts) into a single figure using **subplots**. This is useful when you want to display different types of visualizations side-by-side for comparative purposes or for a more comprehensive view of the data.

1. Using Subplots in Matplotlib

Subplots allow you to arrange multiple plots in a grid layout. You can specify the number of rows and columns in the grid, and then plot different visualizations in each grid cell.

2. Example: Combining Line Chart, Bar Chart, and Pie Chart in Subplots

In this example, we'll create a figure that contains three different plots:

A **line chart** showing a trend over time.

A **bar chart** representing categorical data.

A **pie chart** showing the proportions of categories.

Code Example:

```
import matplotlib.pyplot as plt
import numpy as np

# Sample data
x = np.arange(1, 6)
y1 = [2, 4, 6, 8, 10] # Line chart data
y2 = [5, 3, 6, 2, 7] # Bar chart data
labels = ['A', 'B', 'C', 'D', 'E'] # Pie chart categories
sizes = [15, 30, 45, 10, 20] # Pie chart data

# Create a figure with 1 row and 3 columns
fig, axs = plt.subplots(1, 3, figsize=(15, 5))

# Line chart in the first subplot
axs[0].plot(x, y1, marker='o', color='b', label='Trend')
axs[0].set_title('Line Chart')
axs[0].set_xlabel('X Axis')
axs[0].set_ylabel('Y Axis')
axs[0].legend()

# Bar chart in the second subplot
axs[1].bar(x, y2, color='g', label='Values')
axs[1].set_title('Bar Chart')
axs[1].set_xlabel('Categories')
axs[1].set_ylabel('Values')
axs[1].set_xticks(x)
axs[1].set_xticklabels(labels)
```

```
axs[1].legend()
```

```
# Pie chart in the third subplot
```

```
axs[2].pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)
```

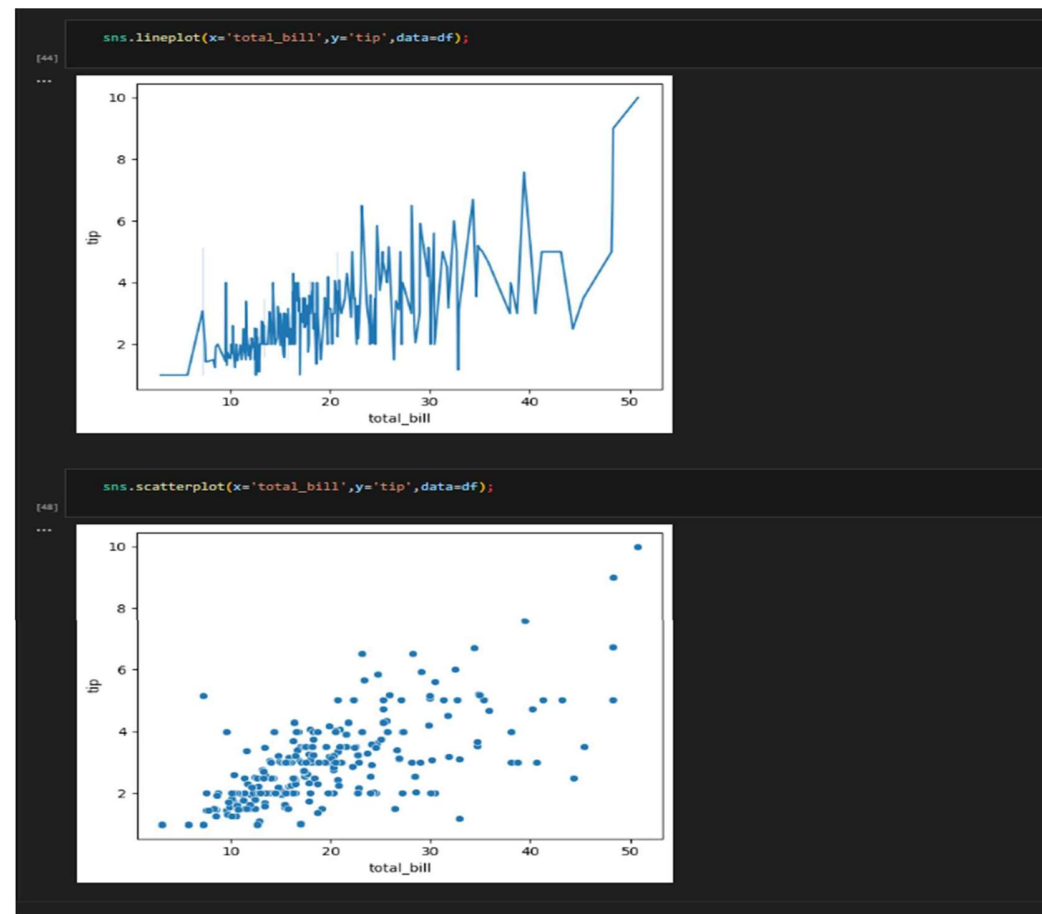
```
axs[2].set_title('Pie Chart')
```

```
# Adjust layout to prevent overlap
```

```
plt.tight_layout()
```

```
# Show the plot
```

```
plt.show()
```



Topic: Summary of Key Learnings

- Documented techniques learned over the past weeks.
- Example: Listed best practices for data cleaning and visualization.

1. Data Cleaning Techniques**Handling Missing Data:**

Imputation: Filling missing values using mean, median, or mode (for numerical data) or the most frequent value (for categorical data).

Removal: Dropping rows or columns with too many missing values.

Interpolation: For time series or sequential data, missing values can be interpolated based on surrounding data points.

Example:

```
df.fillna(df.mean(), inplace=True) # Impute missing values with column mean
```

Data Transformation:

Normalization/Standardization: Scaling numeric data to a standard range, often required for machine learning models.

Log Transformation: Used to deal with skewed distributions by applying a logarithmic scale.

Categorical Encoding: Converting categorical variables into numeric formats using one-hot encoding or label encoding.

Example:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df['scaled_column'] = scaler.fit_transform(df[['column']])
```

Outlier Detection and Removal:

Z-Score Method: Identifying and removing data points that deviate significantly from the mean (e.g., z-scores greater than 3).

IQR Method: Removing data points outside the interquartile range ($Q1 - 1.5 * IQR$, $Q3 + 1.5 * IQR$).

Example:

```
from scipy import stats
df = df[(np.abs(stats.zscore(df['column'])) < 3)] # Remove outliers based on Z-score
```

2. Combining Multiple Datasets

Concatenation: Combining datasets vertically (stacking rows) or horizontally (adding columns) using `concat()`.

Merging: Joining datasets based on common columns or indices using `merge()` (similar to SQL joins).

30-11-2024

Training Day – 50

Topic: Final Review and Practice

- Revisited core concepts and practiced integrating analysis and visualization.
- Example: Created a summary report of the entire analysis workflow.

The screenshot shows a Jupyter Notebook interface with the following content:

Airline Dataset

Importing Required Modules

1. importing numpy for mathematical operation on arrays and dataframe.
2. importing pandas for reading data and data manipulation.
3. importing matplotlib and seaborn to show the insights and visualization from the dataset.
4. importing warnings for Warning messages that are typically issued in dataframe where it is useful to alert the user of some condition in a program, where that condition (normally) doesn't warrant raising an exception and term

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

107

108

109

110

111

112

Reading Dataset and Checking the NaN Values , Data Types , and Statistical Analysis

1. Since data is in form of excel file we have to use pandas read_excel to load the data
2. After loading it is important to check the complete information of data as it can indicate many of the hidden information such as null values in a column or a row
3. Check whether any null values are there or not, if it is present then following can be done.
 1. Filling NaN values with mean, median and mode using fillna() method
4. Describe data --> which can give statistical analysis

```
df=pd.read_excel("Data_train.xlsx")
```

118

```
df.head()
```

119

120

121

```
df.shape
```

122

123

124

125

```
df.info()
```

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511

1512

1513

1514

1515

1516

1517

15

```
Pandas-Exercises-Basic-Understanding.ipynb | Copy of Copy of Copy of Pandas-Exercises-Apple-Stock (1) (1) (1).ipynb | Pandas-Exercises-Aggregation.ipynb | Pokemon (1) (1) (4) (1).ipynb | AIRL...
C: > Users > Roshni > Downloads > AIRLINE_FINAL (6) (1) (1).ipynb > ...
+ Code + Markdown | ▶ Run All | Clear All Outputs | Outline ...

df.describe()
[63]
...

df.describe(include=object)
[64]
...

df.isnull().sum()
[65]
...
Airline      0
Date_of_Journey  0
Source        0
Destination   0
Route         1
Dep_Time      0
Arrival_Time  0
Duration      0
Total_Stops   1
Additional_Info 0
Price         0
dtype: int64

df['Route'].mode()
[66]
...
0 DEL → BOM → COK
Name: Route, dtype: object

df['Route']=df['Route'].fillna(df['Route'].mode()[0])
[67]

df['Total_Stops'].mode()
[68]
...
0 1 stop
Name: Total_Stops, dtype: object

df['Total_Stops']=df['Total_Stops'].fillna(df['Total_Stops'].mode()[0])
```

```
Pandas-Exercises-Basic-Understanding.ipynb | Copy of Copy of Copy of Pandas-Exercises-Apple-Stock (1) (1) (1).ipynb | Pandas-Exercises-Aggregation.ipynb | Pokemon (1) (1) (4) (1).ipynb | AIRLINE_FINAL

+ Code | + Markdown | ▶ Run All | ⌵ Clear All Outputs | ⌵ Outline | ...

From df.info() we can see that Date_of_Journey is a object data type

1. Therefore, we have to convert this datatype into timestamp so that we can use that column properly to find the insights.
2. For this we require pandas to_datetime to convert object data type to datetime dtype.

df['Date_of_Journey']=pd.to_datetime(df['Date_of_Journey'])

df.head(2)

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 11 columns):
#   Column              Non-Null Count  Dtype  
---  --
0   Airline              10683 non-null  object  
1   Date_of_Journey      10683 non-null  datetime64[ns]
2   Source              10683 non-null  object  
3   Destination          10683 non-null  object  
4   Route               10683 non-null  object  
5   Dep_Time            10683 non-null  object  
6   Arrival_Time        10683 non-null  object  
7   Duration            10683 non-null  object  
8   Total_Stops          10683 non-null  object  
9   Additional_Info      10683 non-null  object  
10  Price               10683 non-null  int64   
dtypes: datetime64[ns](1), int64(1), object(9)
memory usage: 918.2+ KB

df['Total_Stops'].unique()

array(['non-stop', '2 stops', '1 stop', '3 stops', '4 stops'],
      dtype=object)
```

1st Insights: How many Flights with respect to their Stopages ?

```
df['Total_Stops'].value_counts()
```

```
Total_Stops
1    5626
0    3491
2    1520
3     45
4      1
Name: count, dtype: int64
```

```
# From This Histogram we can see that no. of flights and their Stopages
# In this Data maximum flights have 1 stopages
# And there are few flights which have 3rd and 4th stopages
```

```
plt.title('No. of flights Stopage')
plt.hist(df['Total_Stops'], color='purple')
plt.xlabel('Stopages')
plt.ylabel('No. of Flights')
plt.xticks(df['Total_Stops'].unique())
plt.show()
```

