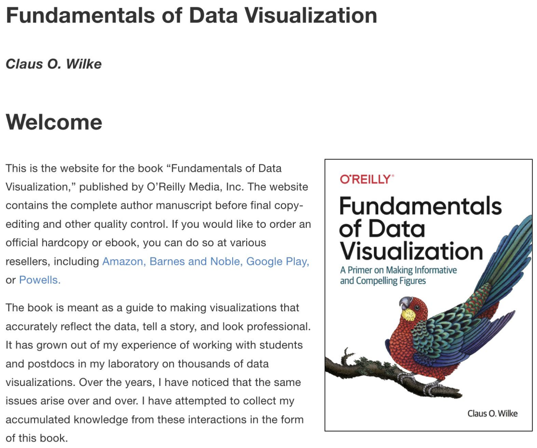
SICSS Exploratory Data Analysis and Visualization

Dr. Uzay Çetin



Fundamentals of Data Visualization

Claus O. Wilke

Welcome

This is the website for the book “Fundamentals of Data Visualization,” published by O’Reilly Media, Inc. The website contains the complete author manuscript before final copy-editing and other quality control. If you would like to order an official hardcopy or ebook, you can do so at various resellers, including Amazon, Barnes and Noble, Google Play, or Powells.

The book is meant as a guide to making visualizations that accurately reflect the data, tell a story, and look professional. It has grown out of my experience of working with students and postdocs in my laboratory on thousands of data visualizations. Over the years, I have noticed that the same issues arise over and over. I have attempted to collect my accumulated knowledge from these interactions in the form of this book.

O'REILLY
Fundamentals of Data Visualization
A Primer on Making Informative and Compelling Figures
Claus O. Wilke

Why learn Data Visualization?

- Digital age comes with huge amount of data
    - Need for **making sense of data with visual tools**
- Visual explanation is more effective than other techniques
    - Tell a story and communicate powerfully

Why Data Visualization?

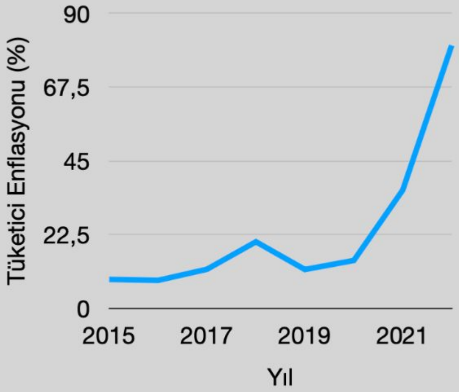
Metin

2015 yılından bu güne Türkiye’deki tüketici enflasyonu incelendiğinde, her yıl bir önceki yıla göre artış göstermiştir. 2015 yılında bir önceki yıla göre %8.81 artan enflasyon izleyen yıllarda sırasıyla, %8.53, %11.92, %20.3, %11.84, %14.6, %36.08 ve %80.21 oranında artış göstermiştir.

Tablo

Yıl	Tüketici Enflasyonu (%)
2015	8.81
2016	8.53
2017	11.92
2018	20.3
2019	11.84
2020	14.6
2021	36.08
2022	80.21

Grafik



Source: ESTU Veri Görselleştirme

Why Data Visualization?

- **A picture is worth a thousand words!** – Frank R. Bernard
  - Psychologist Albert Mehrabian demonstrated that **93% of communication is nonverbal**.
  - Research at 3M Corporation concluded that **we process visuals 60,000 times faster than text**.

Source: Using images Effectively

Why learn Exploratory Data Analysis?

- *The **purpose** of EDA is to use **summary** statistics and visualizations to **better understand data**, and **find clues about the tendencies of the data**, its quality and to **formulate assumptions** and the hypothesis of our analysis. (Source: datascienceguide)*

- EDA is NOT JUST about making FANCY visualizations or even aesthetically pleasing ones
  - The goal is to try and answer questions with data.
  - Create a figure such that it makes you understand the data.
    - get to know the variables and relationships between them.

Why learn Exploratory Data Analysis?

Exploratory Data Analysis (EDA) is a process of describing the data by means of

- statistical
- and visualization techniques

This involves inspecting the dataset from many angles, describing & summarizing it without making any assumptions about its contents.

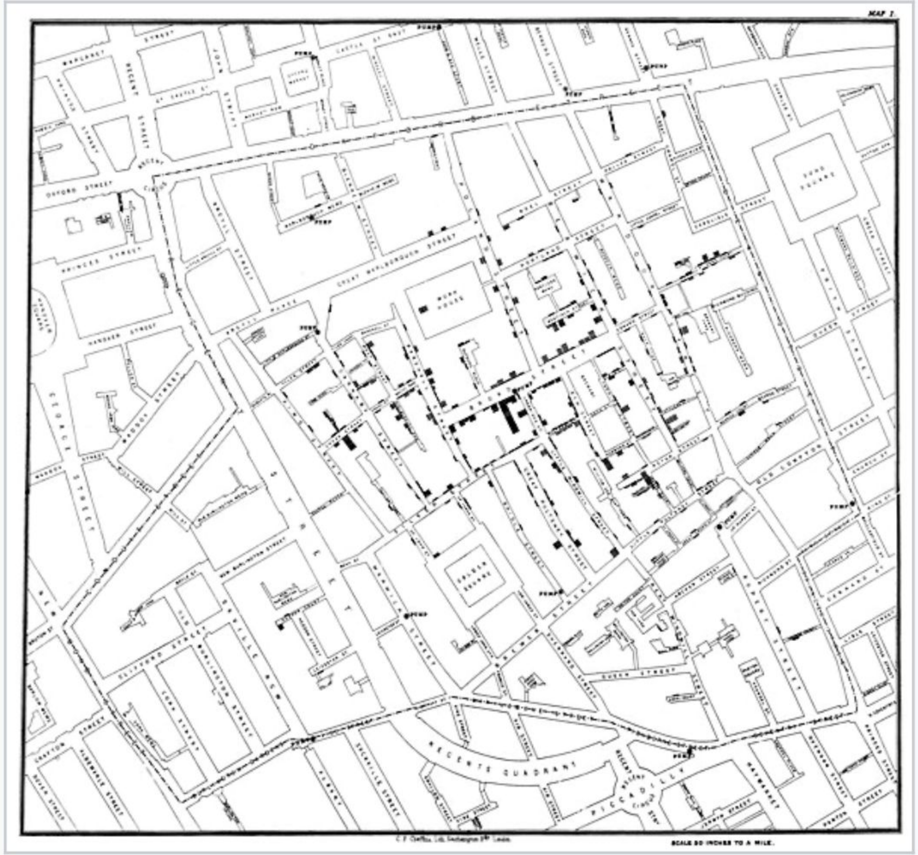
Summary Statistics

Summary statistics are measurements meant to describe data.

- mean, median, mode, max, min, range, quartiles/percentiles, variance, standard deviation, coefficient of determination, skewness and kurtosis.

Why learn Exploratory Data Analysis?

- Understand the nature of the data
  - Snow later used a dot map to illustrate the cluster of cholera cases around the pump. He also used statistics to illustrate the connection between the quality of the water source and cholera cases.

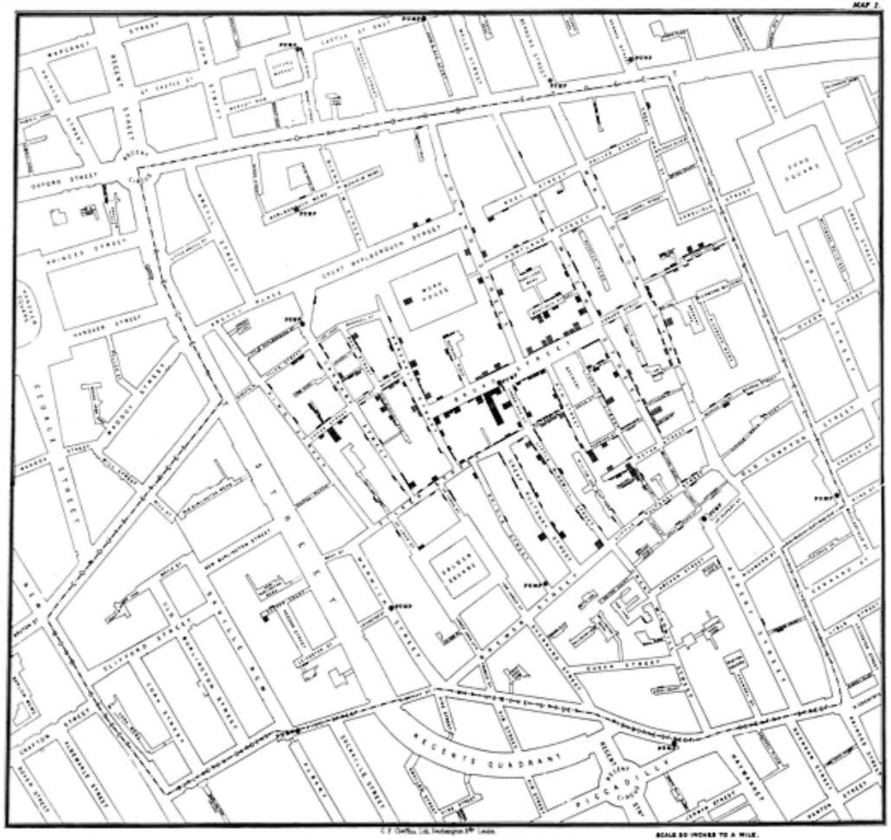


Original map by John Snow showing the clusters of cholera cases in the London epidemic of 1854, drawn and lithographed by Charles Cheffins.

Why learn Exploratory Data Analysis?

Processing data provides a great deal of information. But the million-dollar question is—*how* do we get *meaningful* information from data?

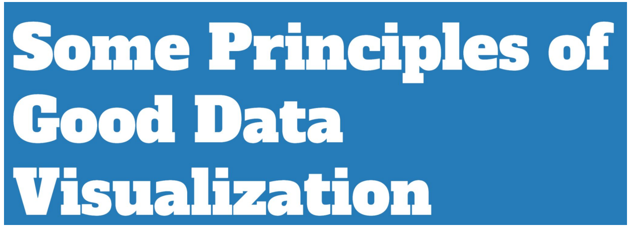
The answer is EDA



Original map by John Snow showing the clusters of cholera cases in the London epidemic of 1854, drawn and lithographed by Charles Cheffins.

Python programming

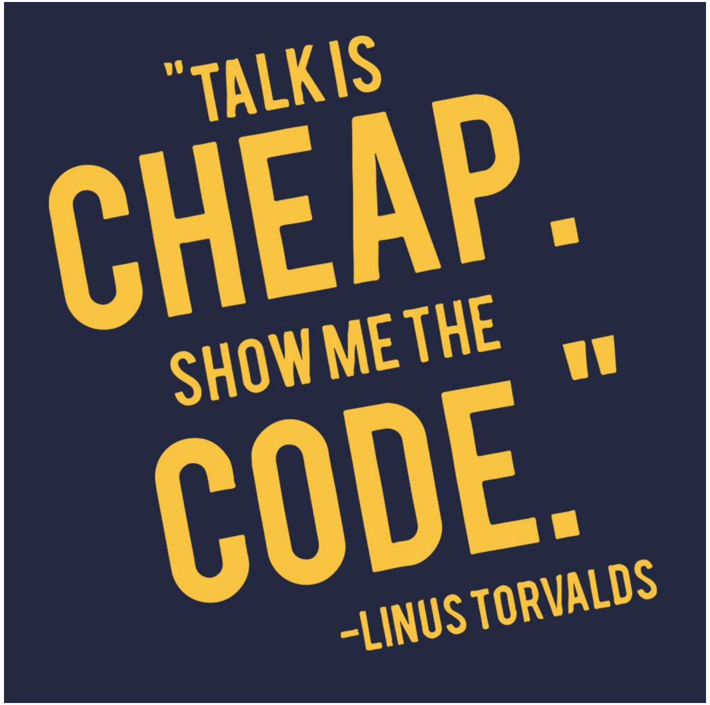
1. No Experience in Coding
2. A little but not yet comfortable
3. Good at coding at least one language but not in Python
4. Comfortable with Python



Some Principles of Good Data Visualization

Source: Data Visualization with Python

Lets Code



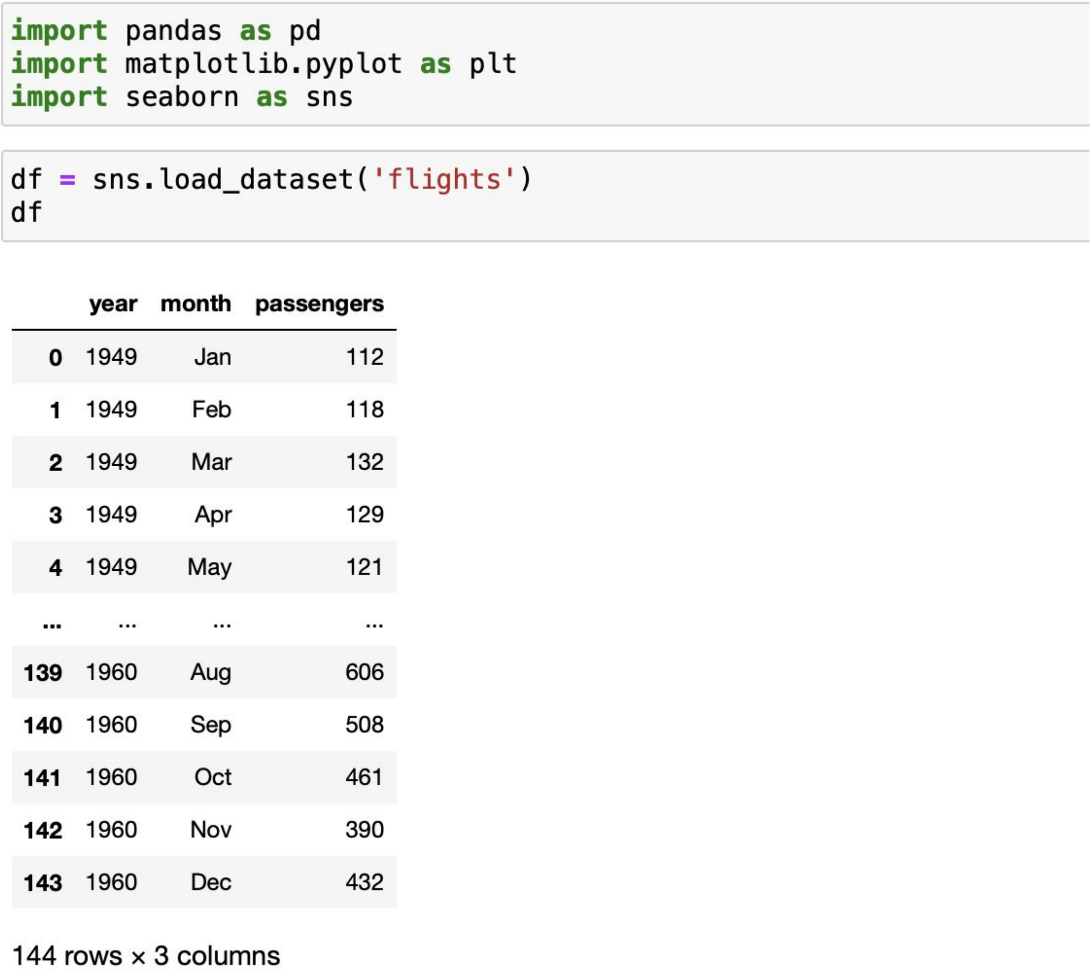
"TALK IS CHEAP. SHOW ME THE CODE."

-LINUS TORVALDS

Source: Data visualization with Python

Lets Code

Open Source Data



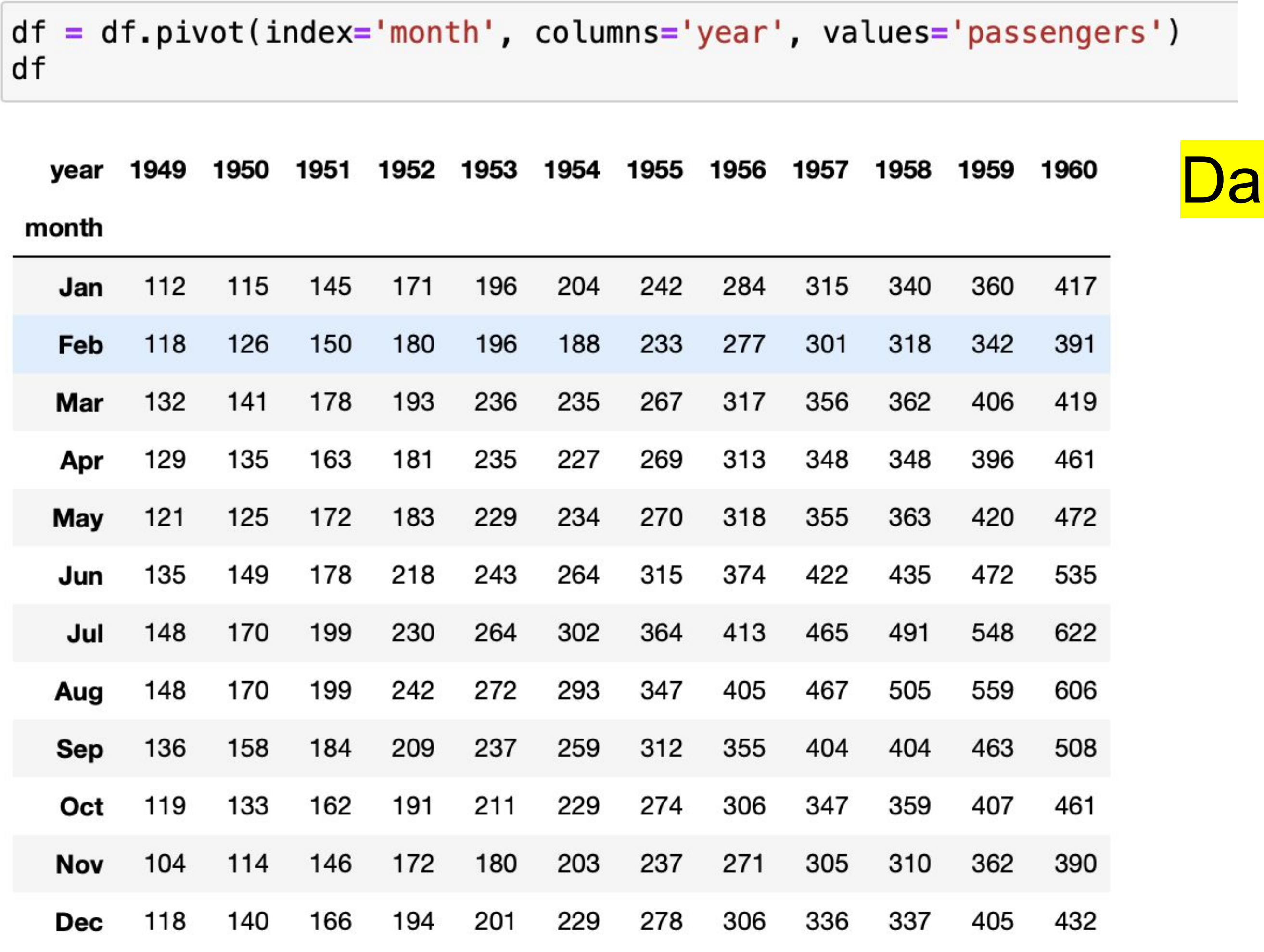
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
df = sns.load_dataset('flights')
df
```

	year	month	passengers
0	1949	Jan	112
1	1949	Feb	118
2	1949	Mar	132
3	1949	Apr	129
4	1949	May	121
...	...	...	...
139	1960	Aug	606
140	1960	Sep	508
141	1960	Oct	461
142	1960	Nov	390
143	1960	Dec	432

144 rows × 3 columns

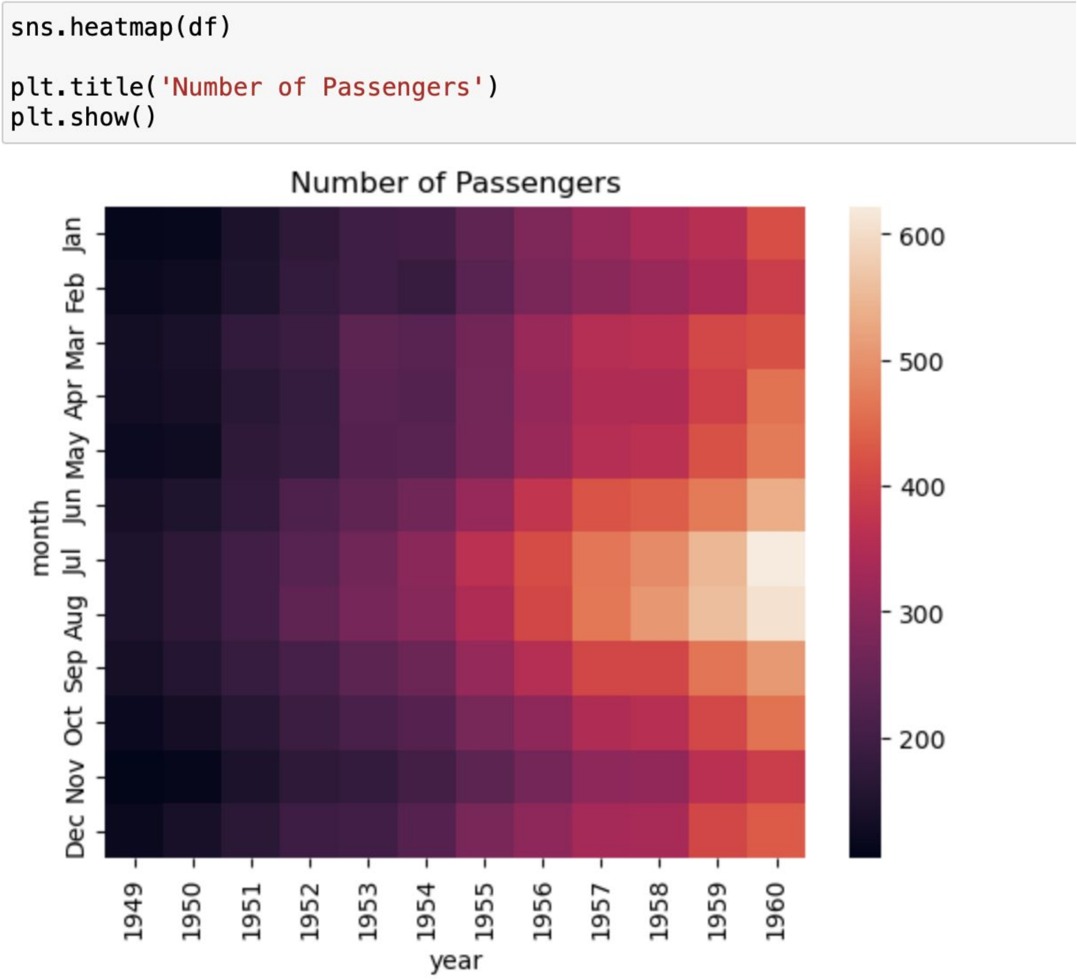
Data Transformation



```
df = df.pivot(index='month', columns='year', values='passengers')
df
```

year	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
month												
Jan	112	115	145	171	196	204	242	284	315	340	360	417
Feb	118	126	150	180	196	188	233	277	301	318	342	391
Mar	132	141	178	193	236	235	267	317	356	362	406	419
Apr	129	135	163	181	235	227	269	313	348	348	396	461
May	121	125	172	183	229	234	270	318	355	363	420	472
Jun	135	149	178	218	243	264	315	374	422	435	472	535
Jul	148	170	199	230	264	302	364	413	465	491	548	622
Aug	148	170	199	242	272	293	347	405	467	505	559	606
Sep	136	158	184	209	237	259	312	355	404	404	463	508
Oct	119	133	162	191	211	229	274	306	347	359	407	461
Nov	104	114	146	172	180	203	237	271	305	310	362	390
Dec	118	140	166	194	201	229	278	306	336	337	405	432

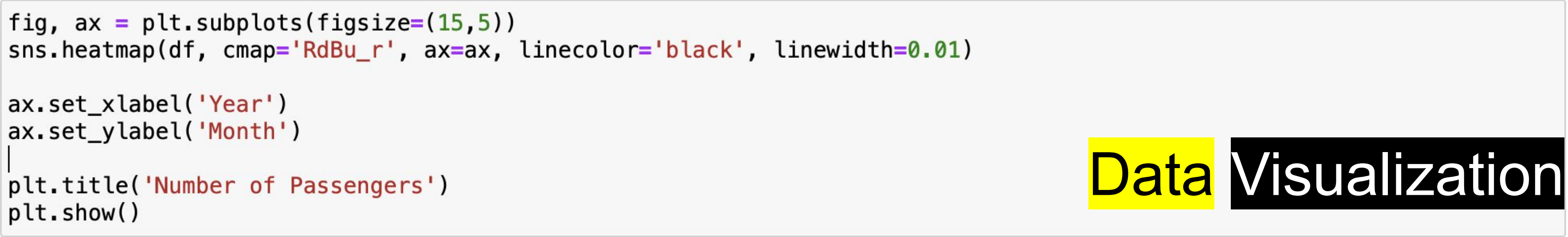
Data Visualization



```
sns.heatmap(df)

plt.title('Number of Passengers')
plt.show()
```

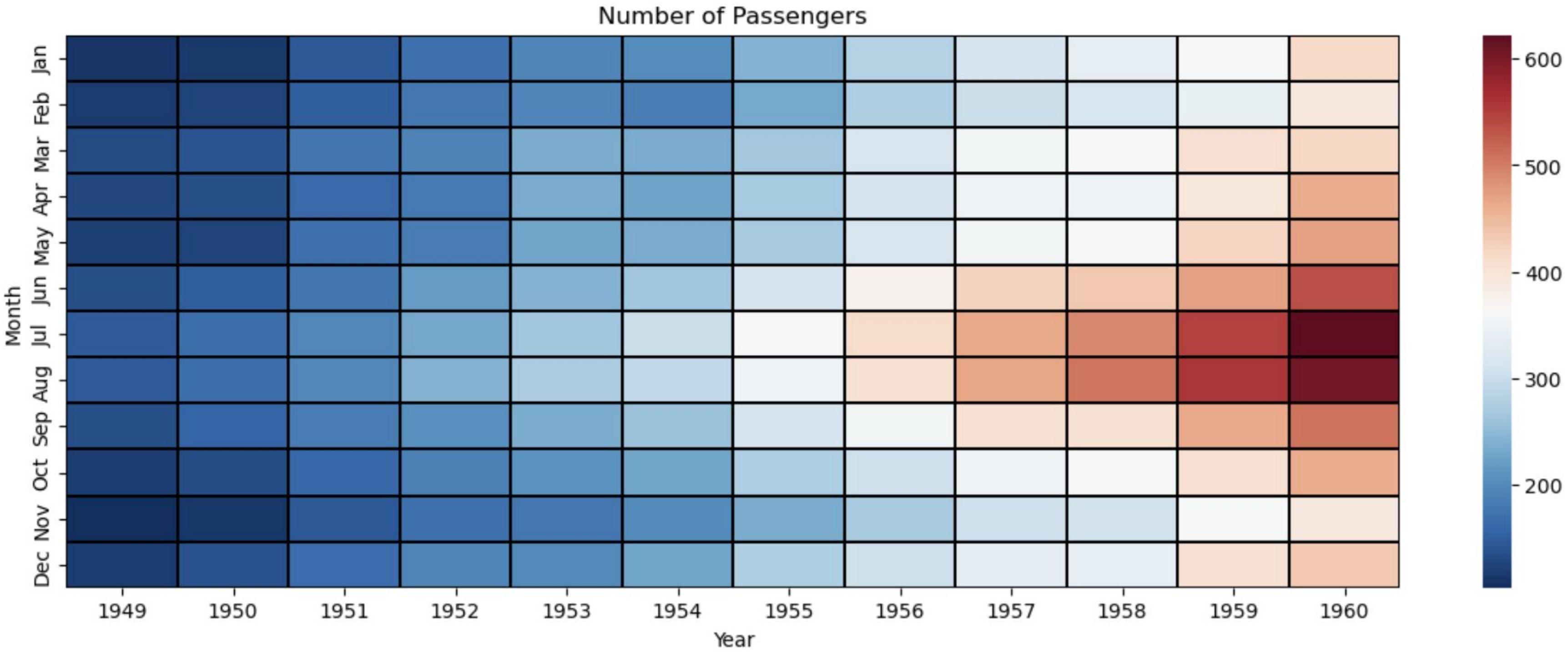
Data Visualization



```
fig, ax = plt.subplots(figsize=(15,5))
sns.heatmap(df, cmap='RdBu_r', ax=ax, linecolor='black', linewidth=0.01)

ax.set_xlabel('Year')
ax.set_ylabel('Month')

plt.title('Number of Passengers')
plt.show()
```

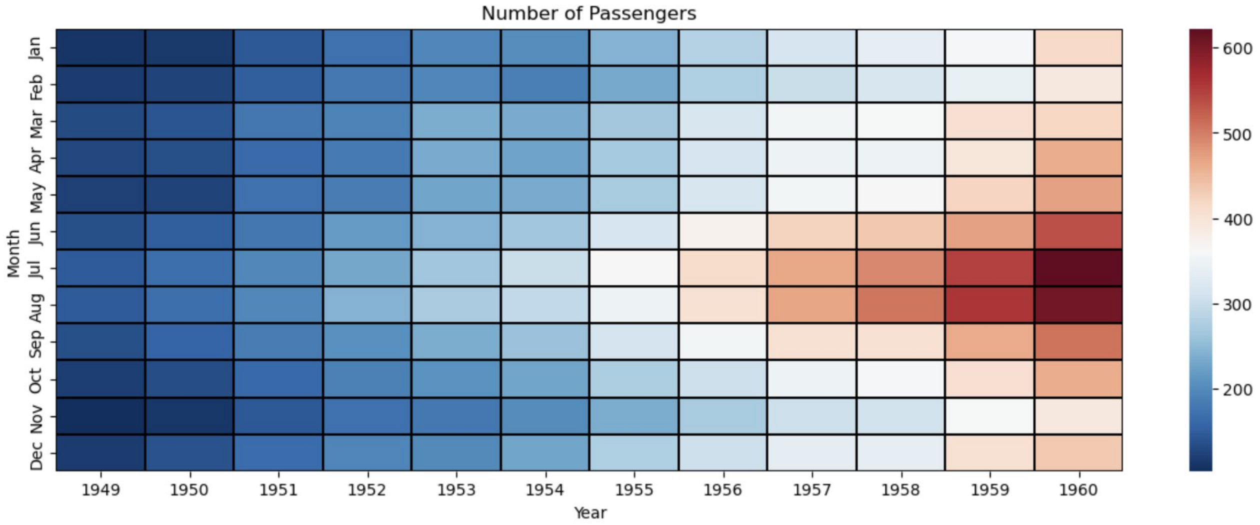


Making Sense of Data

From Data To Information

By looking at the graph, we can infer that :

1. The number of passengers is highest around July and August.
2. The number of passengers grows annually.



A picture is worth a thousand words!

Good Tutorial

Why learn Exploratory Data Analysis?

Raw data is usually skewed, may have outliers, or too many missing values. A model built on such data results in **sub-optimal performance.** (Source: link)

Why learn Exploratory Data Analysis?

- Understand the nature of the data
  - Takes place after feature engineering
  - Done before any modeling

Data Engineer

Data Scientist

Data Analyst

10 minutes to pandas

https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html