Visualization: Python vs R

(matplotlib vs ggplot)

Exploratory Data Analysis and Visualization- Fall 2019- Joyce Robbins

Foad Khoshouei

Nima Chitsazan

Background

Python

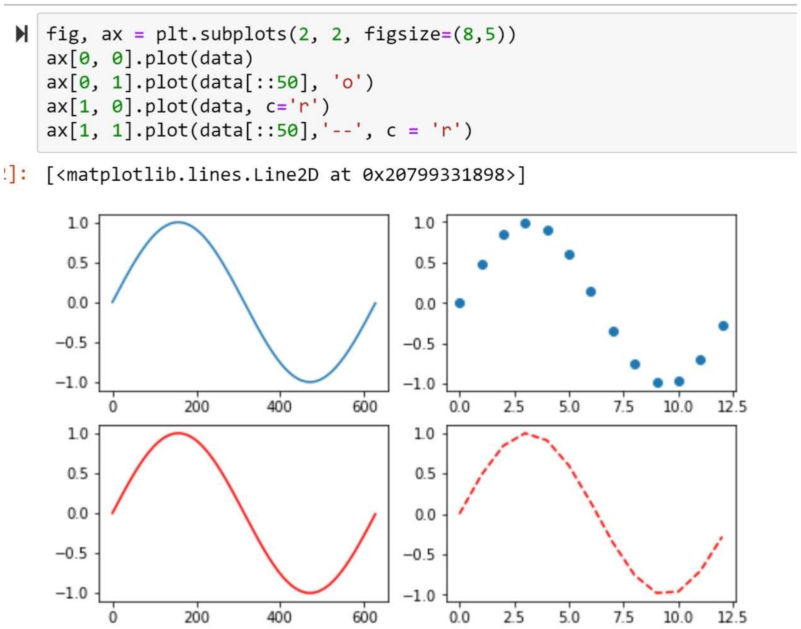
- Started as a hobby project in 1989 by Guido van Rossum
- Python was developed as general-purpose programming language
- Visualization:
  - matplotlib (object oriented)
  - seaborn
  - plotly

R

- Started by Ross Ihaka and Robert Gentleman at University of Auckland in 1992
- Was developed as an open source statistical analysis programming language
- Visualization:
  - ggplot2 (grammar of graphics)
  - lattice
  - plotly

matplotlib

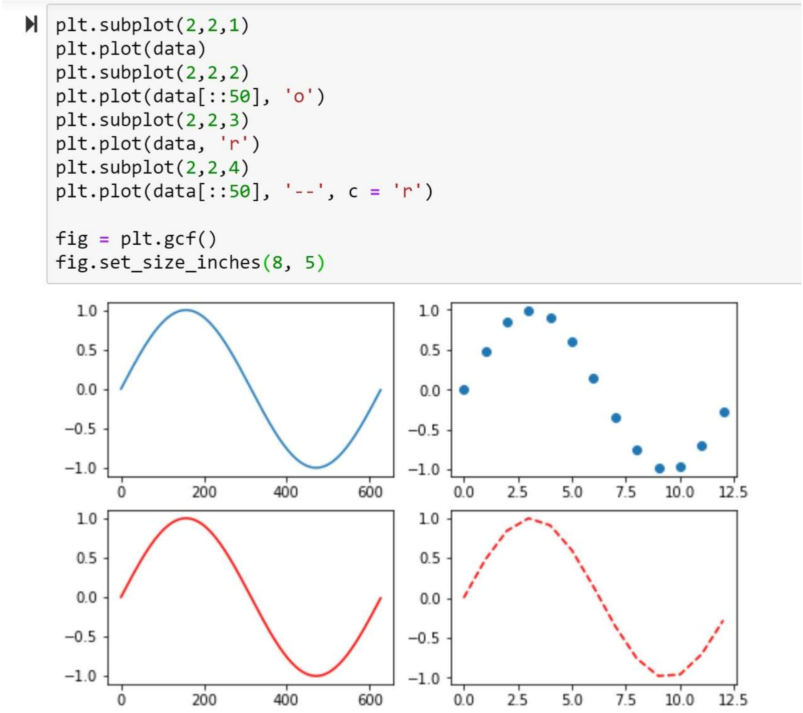
Object-Oriented



```
fig, ax = plt.subplots(2, 2, figsize=(8,5))
ax[0, 0].plot(data)
ax[0, 1].plot(data[::50], 'o')
ax[1, 0].plot(data, c='r')
ax[1, 1].plot(data[::50],'--', c = 'r')
```

```
]: [<matplotlib.lines.Line2D at 0x20799331898>]
```

State-Based

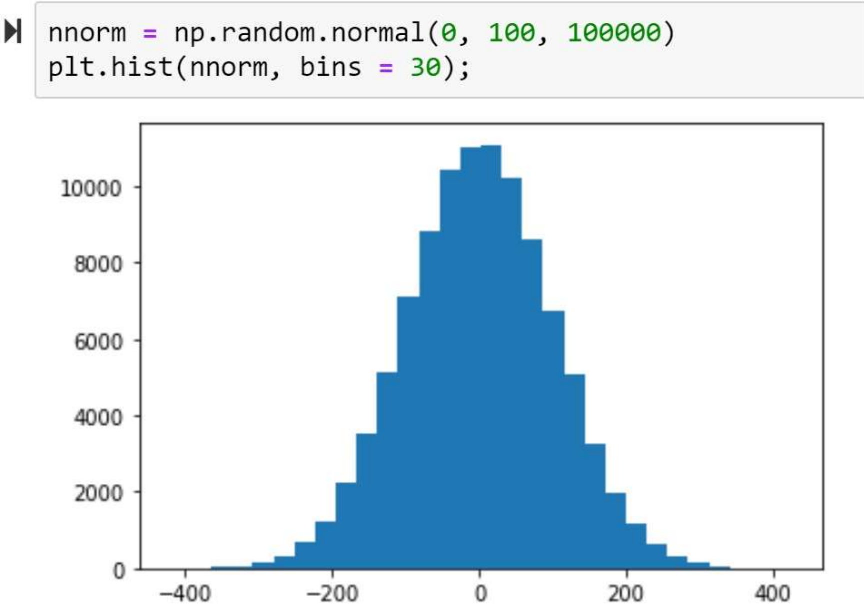


```
plt.subplot(2,2,1)
plt.plot(data)
plt.subplot(2,2,2)
plt.plot(data[::50], 'o')
plt.subplot(2,2,3)
plt.plot(data, 'r')
plt.subplot(2,2,4)
plt.plot(data[::50], '--', c = 'r')

fig = plt.gcf()
fig.set_size_inches(8, 5)
```

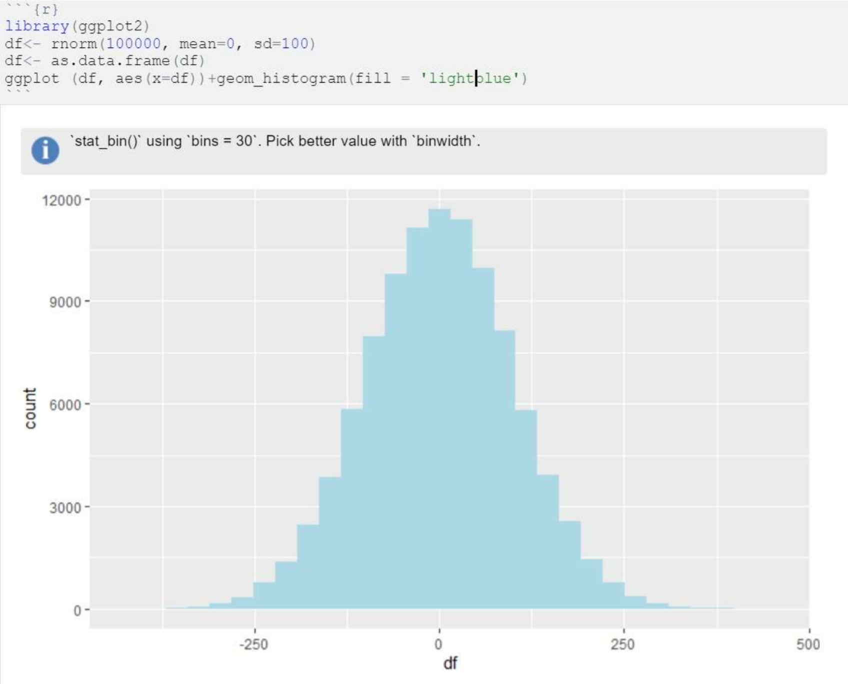
Histogram

python- matplotlib



```python
nnorm = np.random.normal(0, 100, 100000)
plt.hist(nnorm, bins = 30);
```

R- ggplot

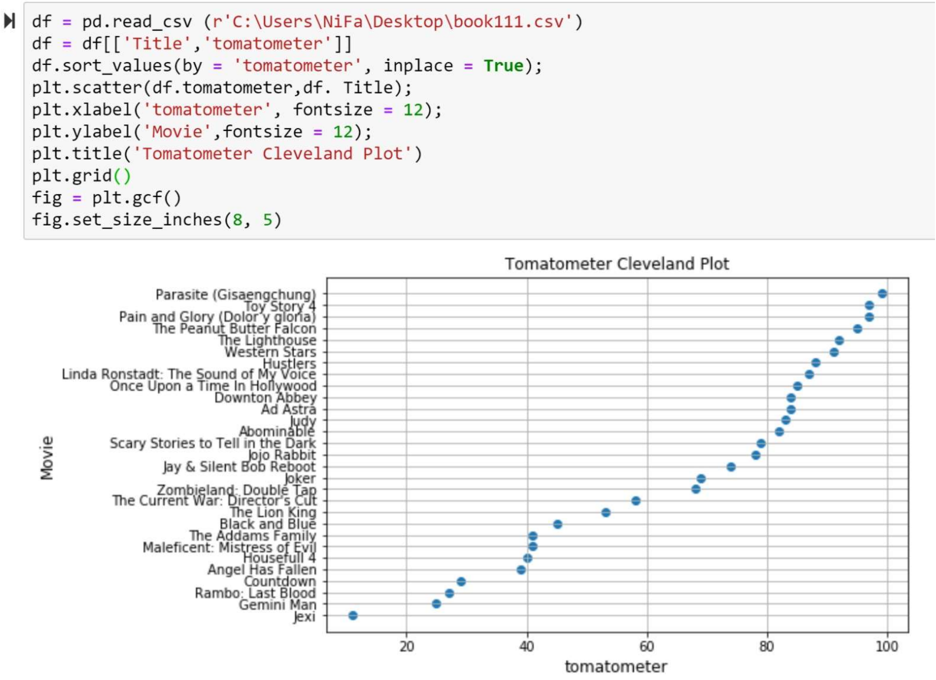


~~~
```{r}
library(ggplot2)
df<- rnorm(100000, mean=0, sd=100)
df<- as.data.frame(df)
ggplot (df, aes(x=df))+geom_histogram(fill = 'lightblue')
```
~~~

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

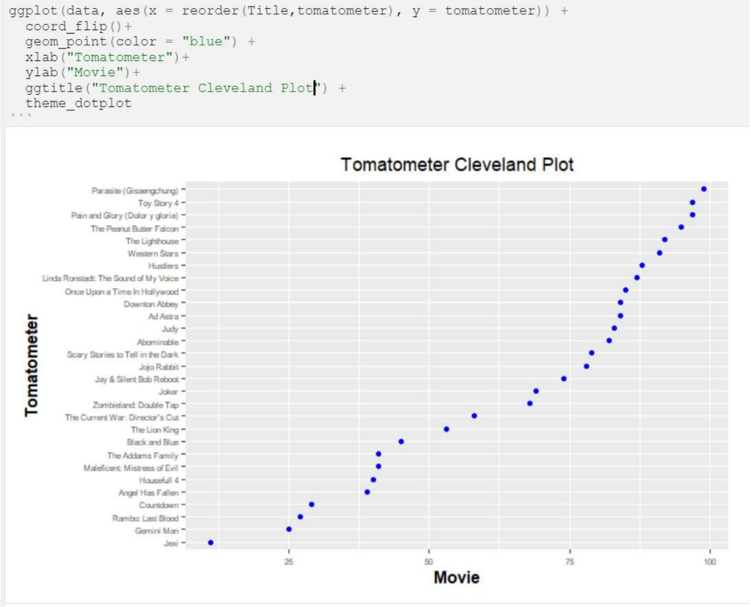
Cleveland Plot

Python- matplotlib



```
df = pd.read_csv (r'C:\Users\NiFa\Desktop\book111.csv')
df = df[['Title','tomatometer']]
df.sort_values(by = 'tomatometer', inplace = True);
plt.scatter(df.tomatometer,df. Title);
plt.xlabel('tomatometer', fontsize = 12);
plt.ylabel('Movie',fontsize = 12);
plt.title('Tomatometer Cleveland Plot')
plt.grid()
fig = plt.gcf()
fig.set_size_inches(8, 5)
```

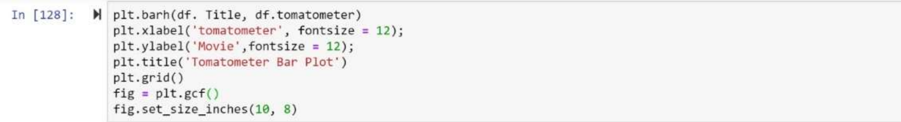
R- ggplot



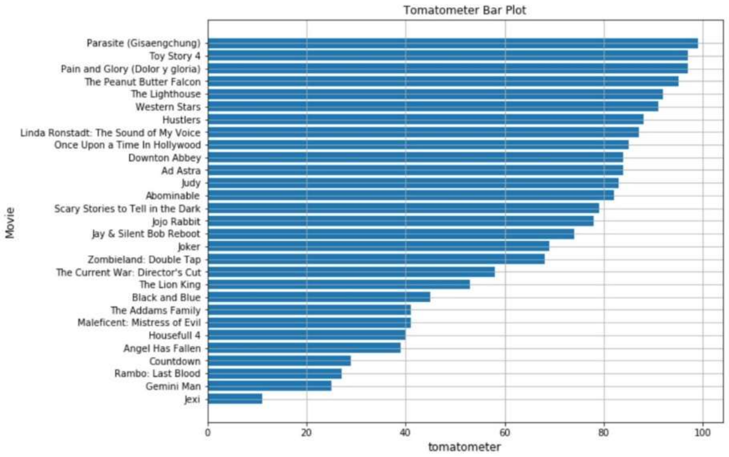
```
ggplot(data, aes(x = reorder(Title,tomatometer), y = tomatometer)) +
  coord_flip()+
  geom_point(color = "blue") +
  xlab("Tomatometer")+
  ylab("Movie")+
  ggtitle("Tomatometer Cleveland Plot") +
  theme_dotplot
```

Bar Plot

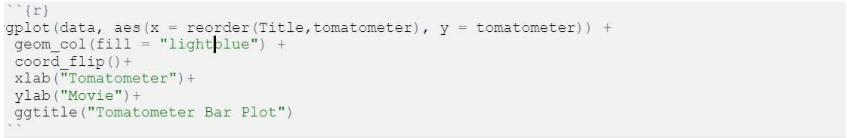
Python- matplotlib



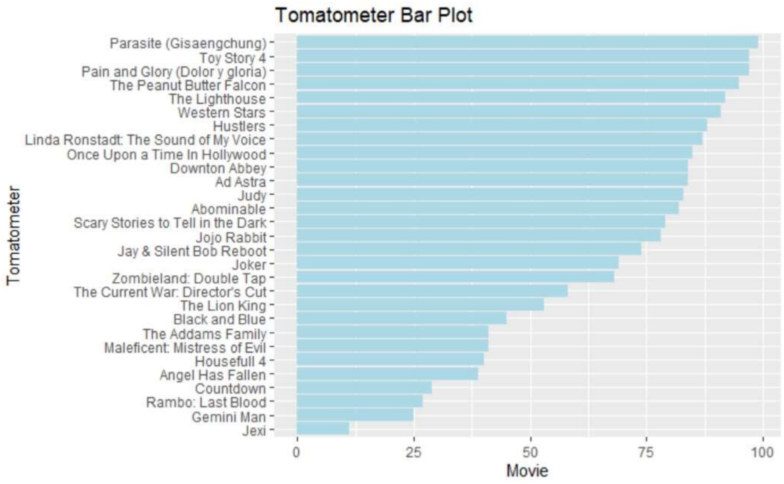
```
In [128]:  plt.barh(df. Title, df.tomatometer)
           plt.xlabel('tomatometer', fontsize = 12);
           plt.ylabel('Movie',fontsize = 12);
           plt.title('Tomatometer Bar Plot')
           plt.grid()
           fig = plt.gcf()
           fig.set_size_inches(10, 8)
```



R- ggplot



~~~
```{r}
gplot(data, aes(x = reorder(Title,tomatometer), y = tomatometer)) +
 geom_col(fill = "lightblue") +
 coord_flip()+
 xlab("Tomatometer")+
 ylab("Movie")+
 ggtitle("Tomatometer Bar Plot")
```
~~~



Summary

- matplotlib: object oriented and state-based
- ggplot uses grammar of graphics
- Seaborn provides high level interface for drawing informative statistical graphics based on matplotlib
- ggplot is more flexible in visualizing complex plots (like mosaic)