Data Science

Intro to Seaborn

CS202 10.03.22

Seaborn

Installation and importing

Just like any other packages we've worked with, seaborn is a package that would require installation. You can type !pip install seaborn on your Jupyter notebook. Or you can type pip install seaborn on your terminal.

Seaborn is a data visualization tool that you can use to produce graphs easily with data frames.

Once you've finished installing you can import to your Jupyter notebook by typing:

```
import sea-born
%matplotlib inline
```

or

```
import seaborn as sns
%matplotlib inline
```

Seaborn

Distribution plots

There are several different types of distribution plots:

- dist plot
- joint plot
- pair plot
- rugplot
- kdeplot

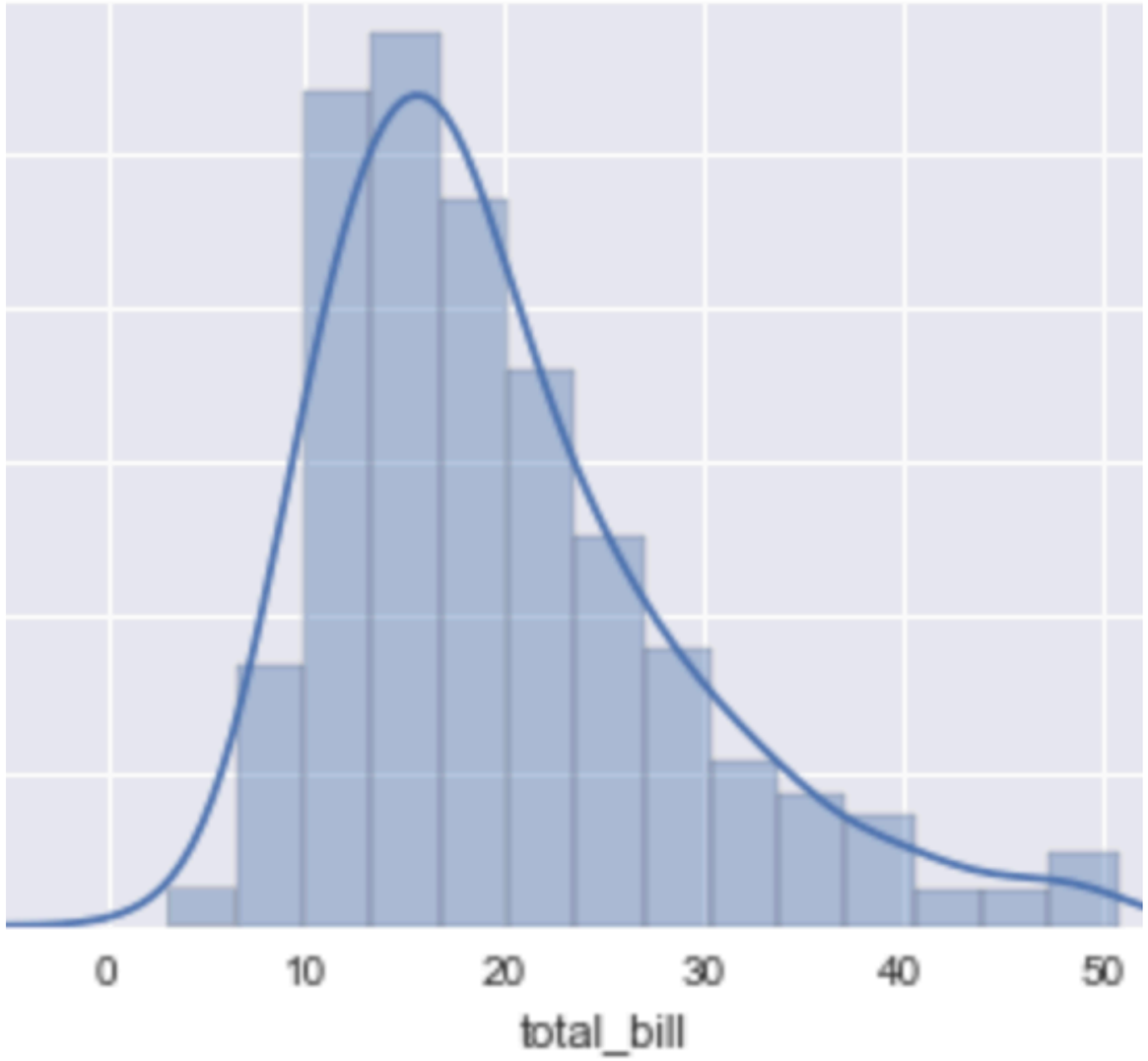
Seaborn comes with example datasets you can use to practice. We will be using one of the built in datasets during the demo. All pictures of graphs in the slides are graphs produced by seaborn using their example dataset.

We will see examples of each in the following slides

Searborn

Distribution plot

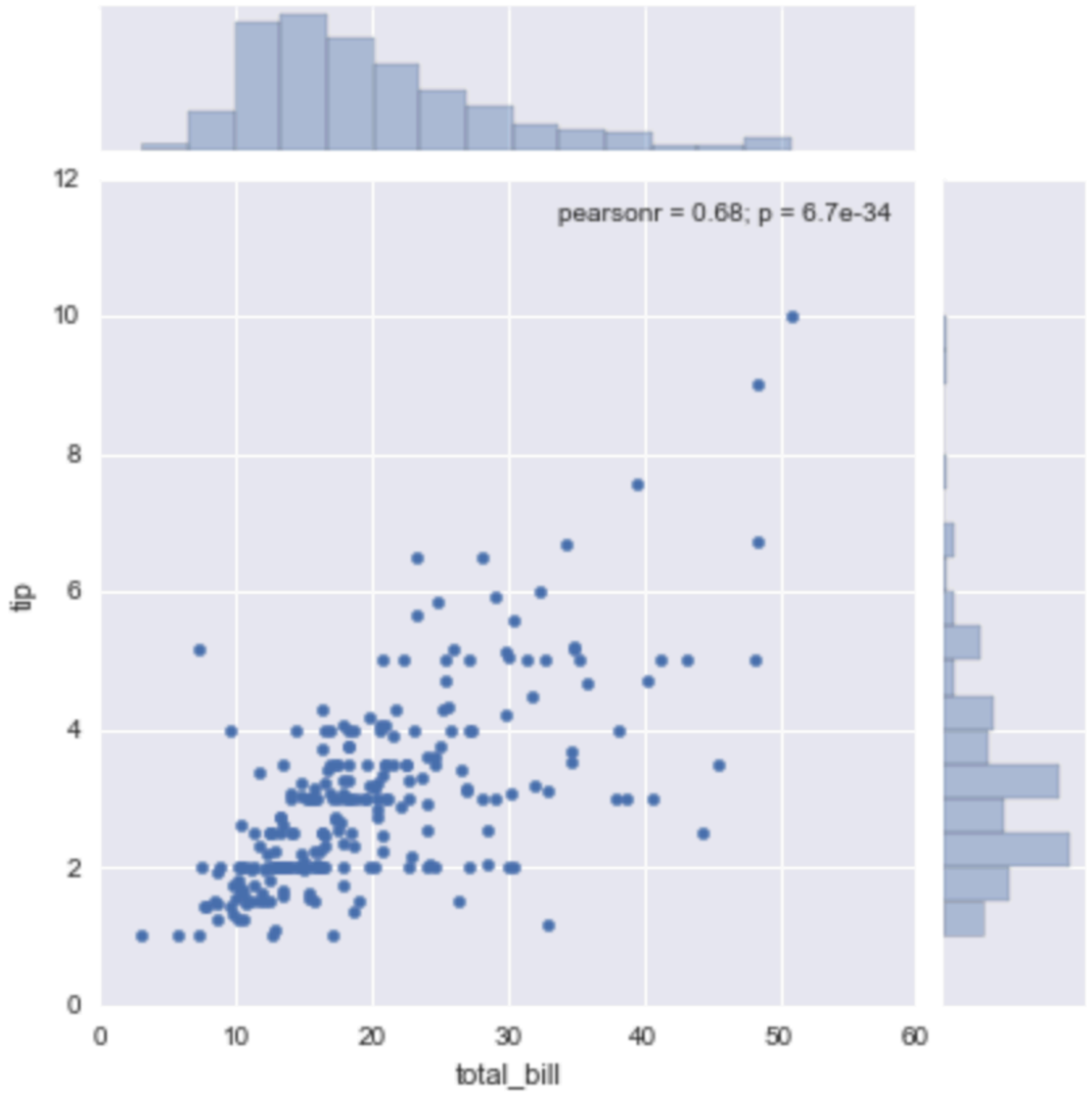
- Distplot
- This looks like your standard bar graph.
- The example on the right is a distribution plot with the kernel density estimation: which is a way to estimate the probability density function of a continuous random variable
- Type the following:
  `sns.distplot(df['col'])`



Seaborn

Distribution plots

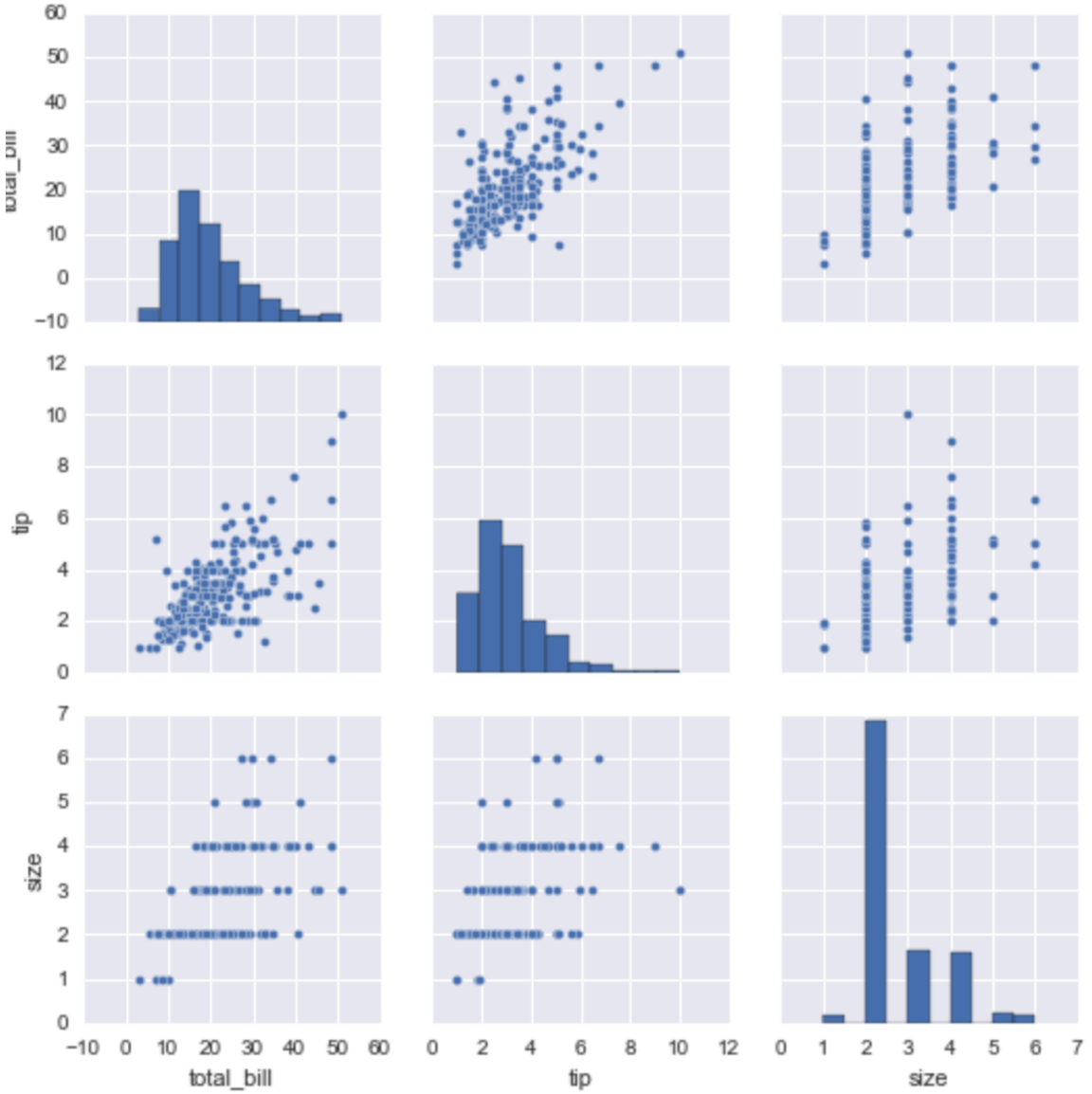
- Joint plot
- A Jointplot comprises three plots. Out of the three, one plot displays a bivariate graph which shows how the dependent variable(Y) varies with the independent variable(X). Another plot is placed horizontally at the top of the bivariate graph and it shows the distribution of the independent variable(X).
- Type the following:
- `sns.jointplot(x = ‘col1’, y=‘col2, data = df, kind = scatter)`



Seaborn

Distribution plots

- Pairplots
- A pairplot plot a pairwise relationships in a dataset.
- The pairplot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column.
- Type the following:
  `sns.pairplot(df)`



Seaborn

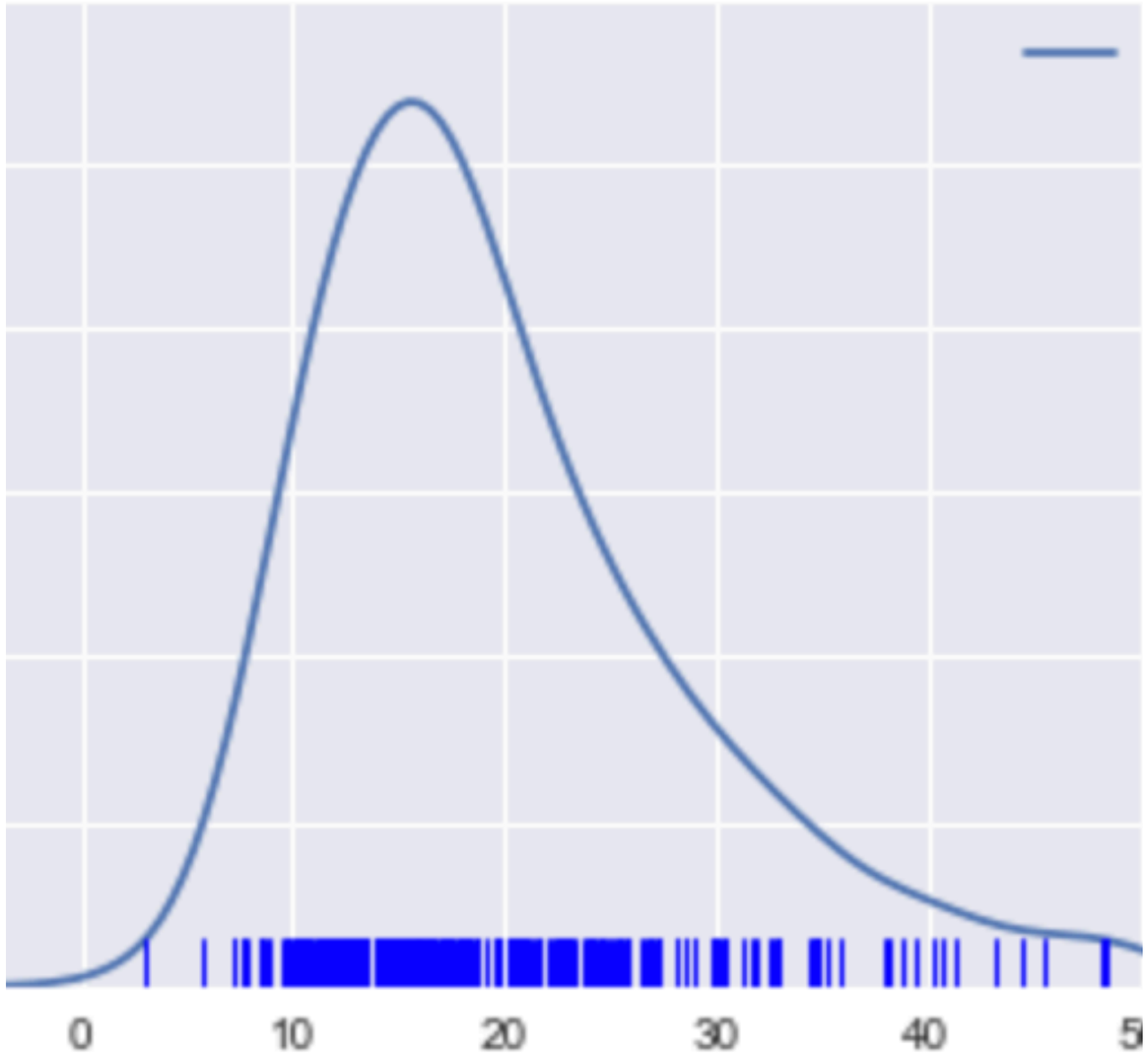
Distribution plots

- Rugplot
- Plot marginal distributions by drawing ticks along the x and y axes.
- This function is intended to complement other plots by showing the location of individual observations in an unobtrusive way.
- rugplots are actually a very simple concept, they just draw a dash mark for every point on a univariate distribution. They are the building block of a KDE plot:
- Type the following:
  `sns.rugplot(df['col'])`

Seaborn

Distribution plots

- Kde plot
- kdeplots are Kernel Density Estimation plots. These KDE plots replace every single observation with a Gaussian (Normal) distribution centered around that value.
- For the simplest type of kde plot one can type the following: `sns.kdeplot(df['col'])`



Seaborn

Categorical plots

There are several different types of categorical plots. Often times in a data set you will have categorical values.

For example: Maybe you are counting the average number of colored jelly beans in a jelly bean bag. Then your data frame may contain a column named “color” and every jelly bean is categorized as “blue”, “red”, “orange”, etc,. These are non-numerical values. You may like to use categorical plots for such values.

These kinds of plots allow us to choose a numerical variable, like age (or color), and plot the distribution of age (or color) for each category in a selected categorical variable.

Seaborn

Categorical plots

There are different types of categorical plots:

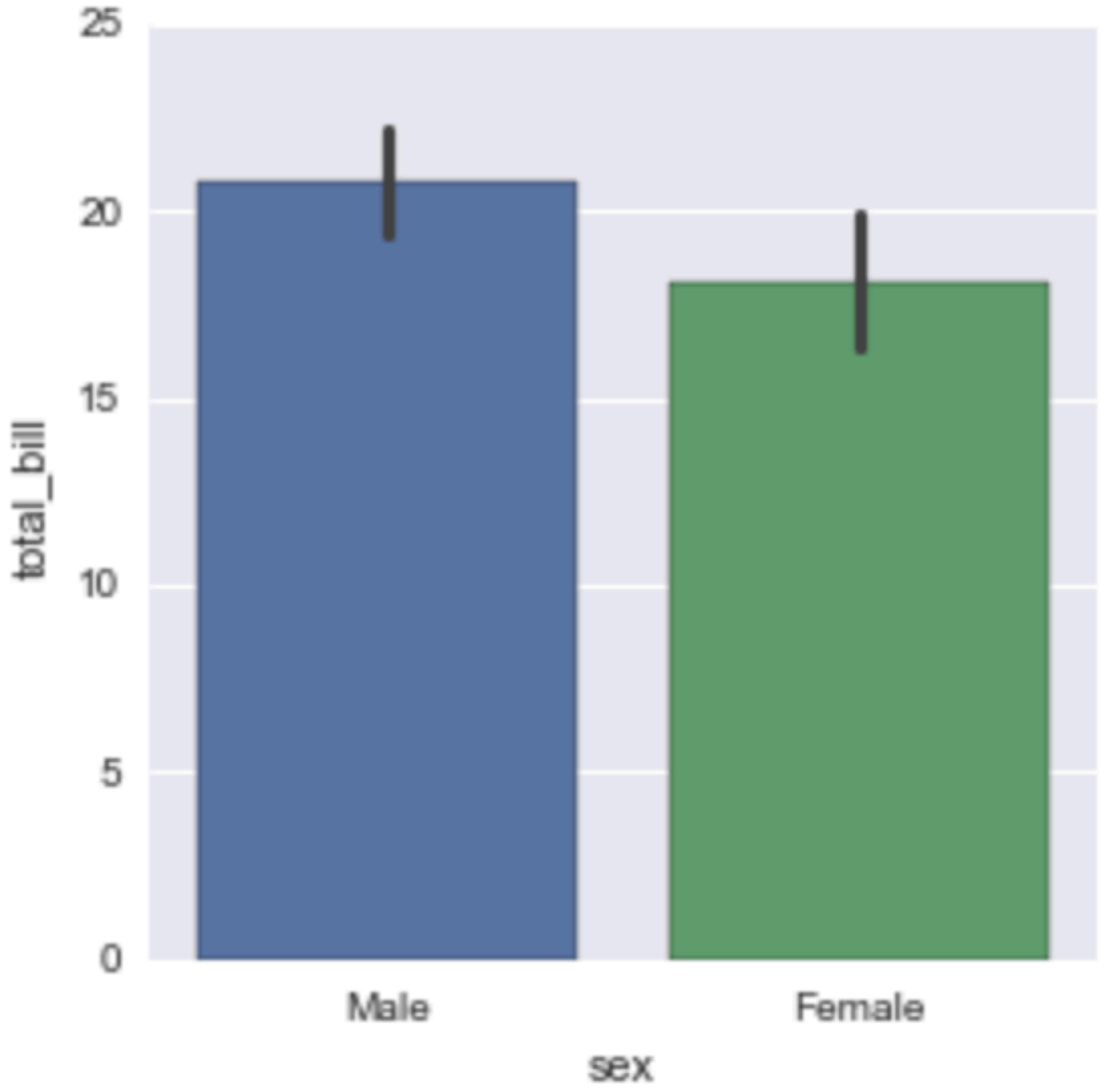
- Factor plot
- box plot
- violin plot
- stripplot
- swarm plot
- bar plot
- count plot

We will see examples of each in the following slides

Seaborn

Categorical plots

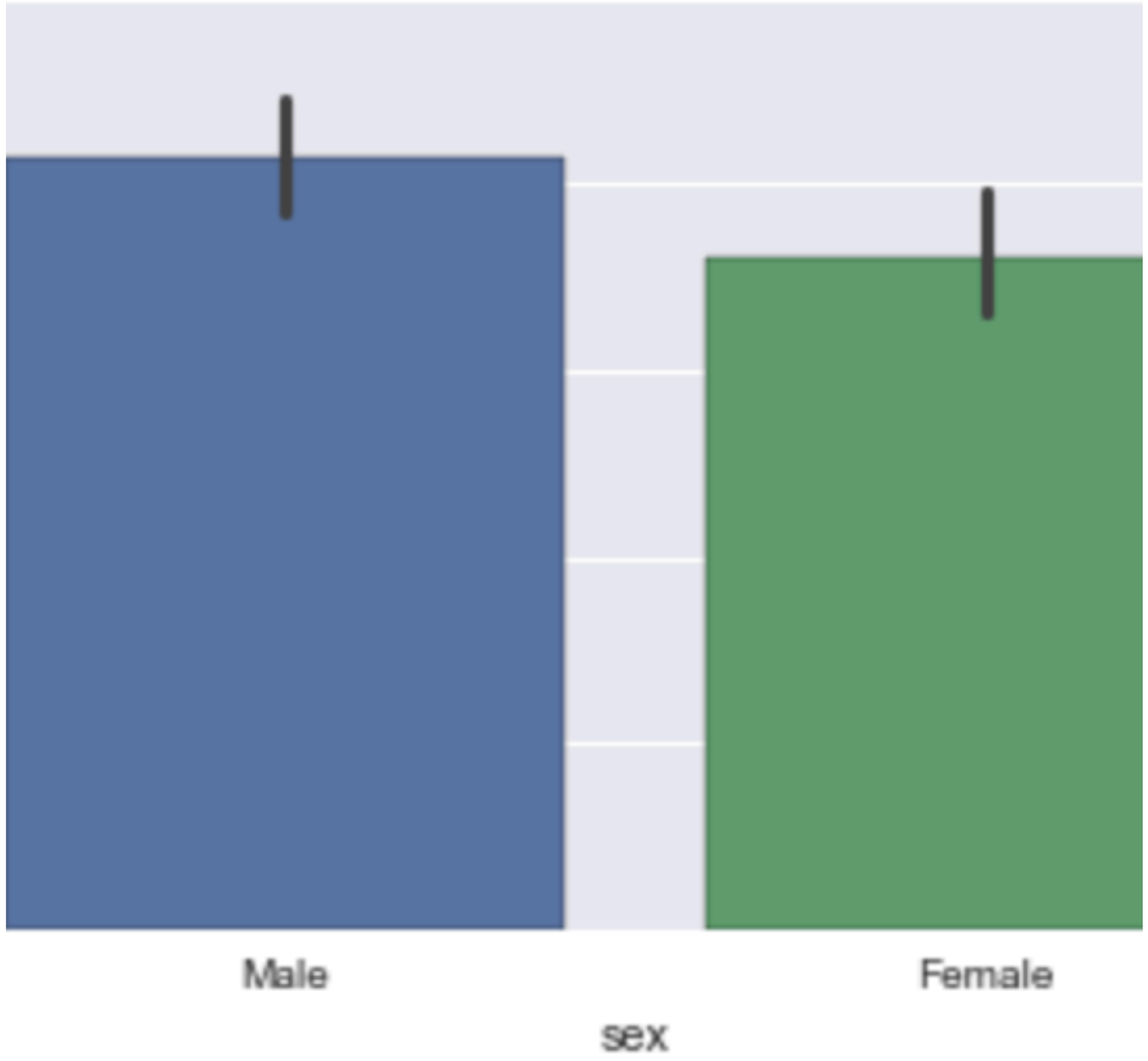
- Factor plot
- Factorplot is the most general form of a categorical plot. It can take in a **kind** parameter to adjust the plot type.
- **`seaborn.factorplot()`** method is used to draw a categorical plot onto a FacetGrid.
- Type the following:
  `sns.factorplot(x = ‘colA’, y = ‘colB’ , data = df, kind = bar)`



Seaborn

Categorical plots

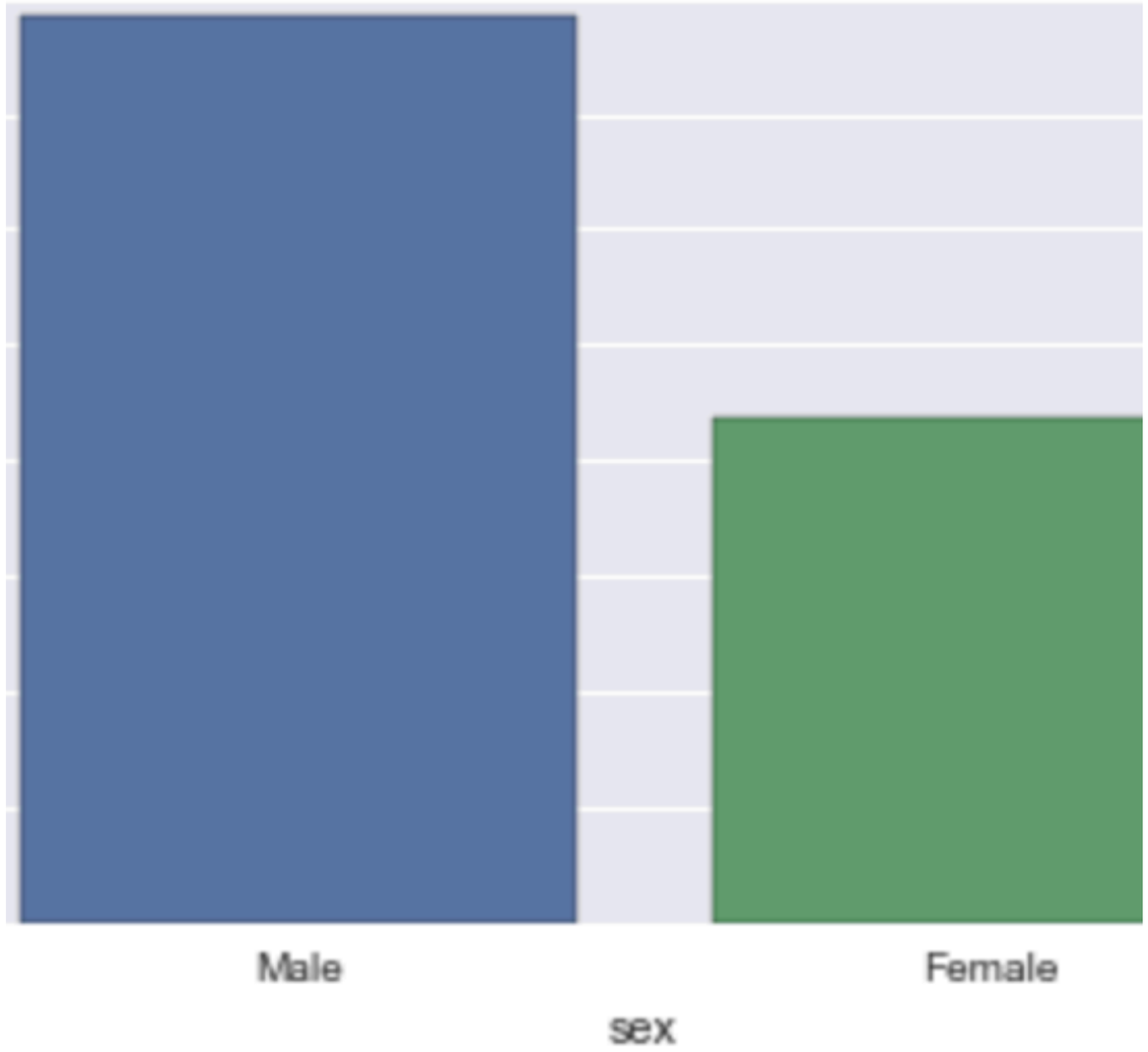
- Bar plot
- These very similar plots allow you to get aggregate data off a categorical feature in your data. **barplot** is a general plot that allows you to aggregate the categorical data based off some function, by default the mean.
- `sns.barplot()` method is used to create a bar plot
- Type the following:
  `sns.barplot(x='col1',y='col2',data=df)`



Seaborn

Categorical plots

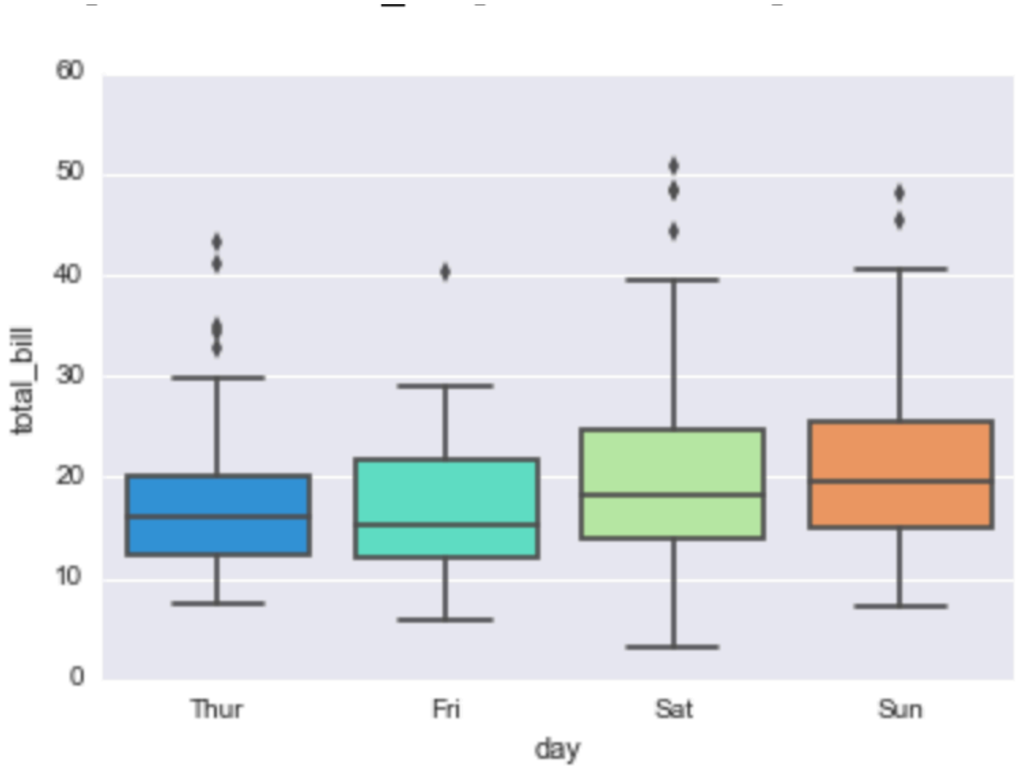
- Countplot
- This is essentially the same as barplot except the estimator is explicitly counting the number of occurrences. Which is why we only pass the x value
- `sns.countplot()` method is used for producing count plots.
- Type the following:
  `sns.countplot(x = ‘col’, data = df)`



Seaborn

Categorical plots

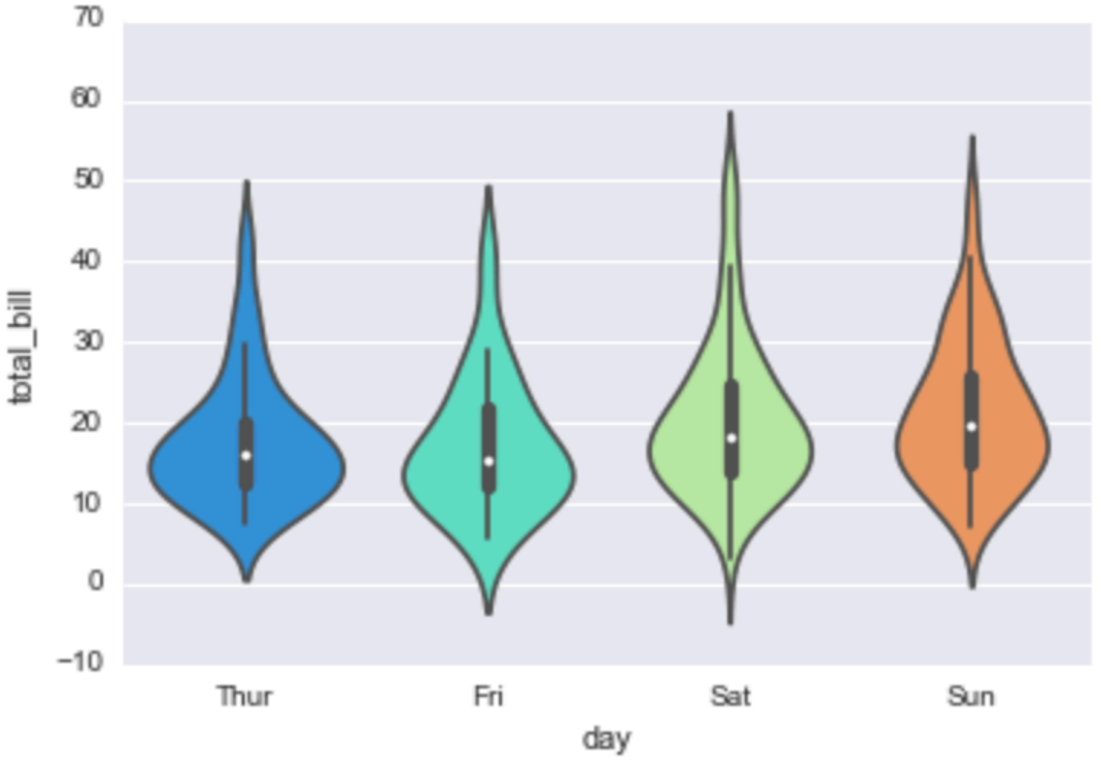
- Boxplot
- boxplots are used to show the distribution of categorical data. A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range.
- sns.boxplot() method is used to produce box plots.
- Type the following:
- `sns.boxplot(x = ‘col1’, y = ‘col2’, data = df)`



Seaborn

Categorical plots

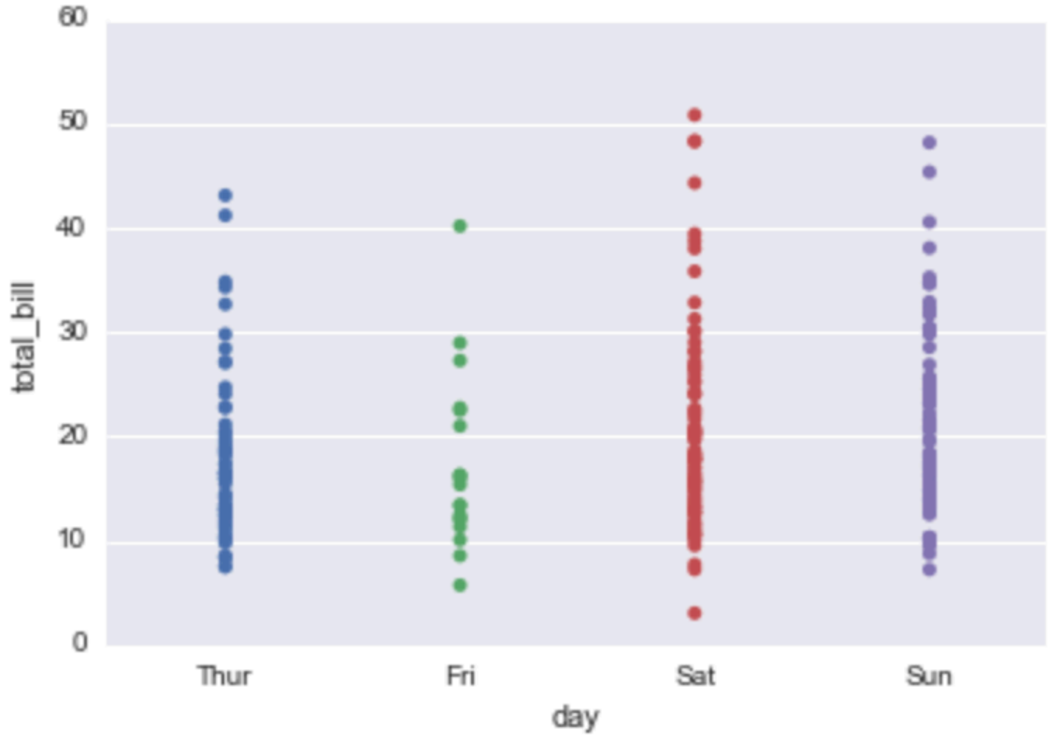
- Violin plot
- A violin plot plays a similar role as a box and whisker plot. It shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared. Unlike a box plot, in which all of the plot components correspond to actual datapoints, the violin plot features a kernel density estimation of the underlying distribution.
- `sns.violinplot()` method is used to plot violin plots
- Type the following:
- `sns.violinplot(x = 'col1', y = 'col2', data = df)`



Seaborn

Categorical plots

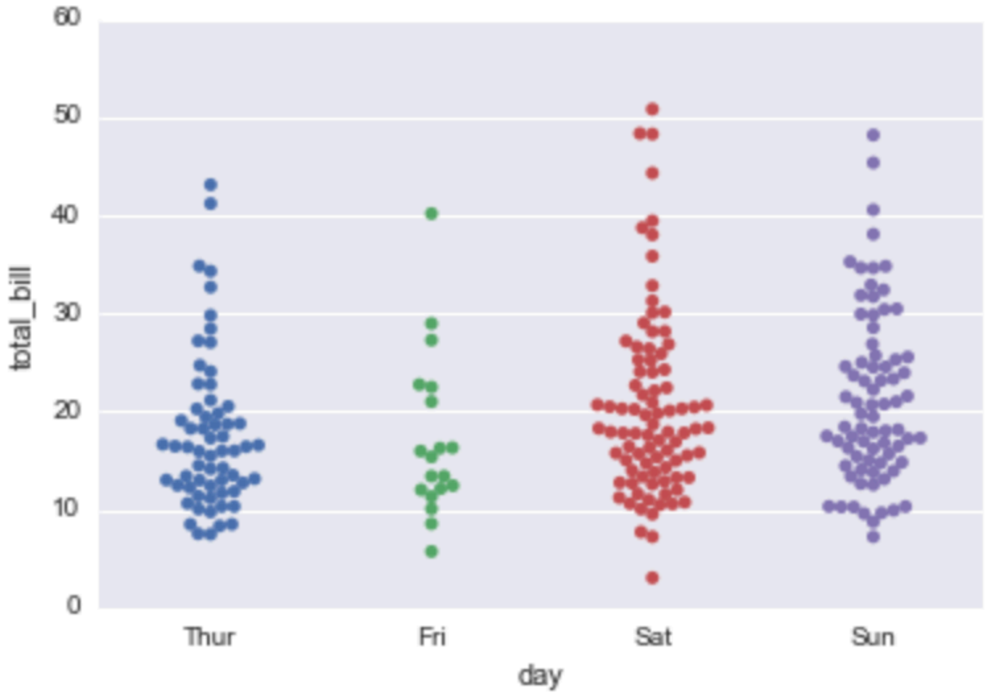
- Strip plot
- The stripplot will draw a scatterplot where one variable is categorical. A strip plot can be drawn on its own, but it is also a good complement to a box or violin plot in cases where you want to show all observations along with some representation of the underlying distribution.
- `sns.stripplot()` method is used to produce strip plots.
- Type the following: `sns.stripplot(x = ‘col1’, y =‘col2’, data = df)`



Seaborn

Categorical plots

- Swarm plot
- The swarmplot is similar to stripplot(), but the points are adjusted (only along the categorical axis) so that they don’t overlap. This gives a better representation of the distribution of values, although it does not scale as well to large numbers of observations (both in terms of the ability to show all the points and in terms of the computation needed to arrange them).
- `sns.swarmplot()` method is used to produce swarm plots
- Type the following: `sns.swarmplot(x = ‘col1’, y =‘col2’, data = df)`



Seaborn

Distribution plots and categorical plots

There are many different ways you can modify your different plots to make them custom to your data and what you are looking for specifically.

We will see at demo some of the modification methods one can use.