Introduction To Matplotlib And Seaborn

DIPESH KUMAR BARAI

M.SC. IN STATISTICS, VISVA BHARATI UNIVERSITY

INDIAN CYBER SECURITY SOLUTIONS

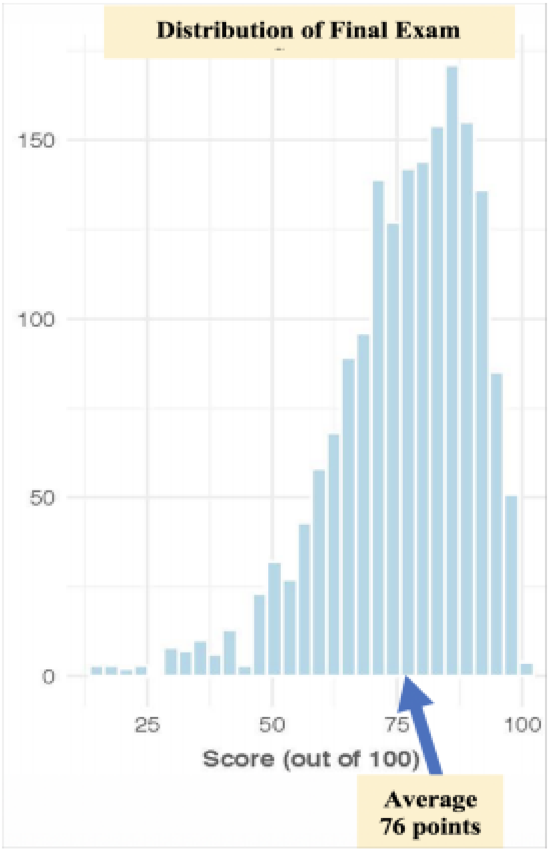
What is Statistics?

- Methodological subject encompassing all aspects of learning from data. tools and methods for working with and understanding data.
- Statisticians apply and develop data analysis methods, seek to understand their properties… …when do these tools provide insight? …when are they possibly misleading?
- Researchers and workers apply and extend statistical methodology,

and contribute new ideas and methods for conducting data analysis.

- A statistic ~ numerical or graphical summary of a collection of data.
- A statistic ~ summary of a collection of data. ○ Average score on

final exam



Variable Types

- Quantitative Variables

Numerical, measurable quantities in which arithmetic operations often make sense.

- Continuous – could take on any value within an interval, many possible values. Ex- Height, Weight.
- Discrete – countable value, finite number of values.(Some set of countable values.)

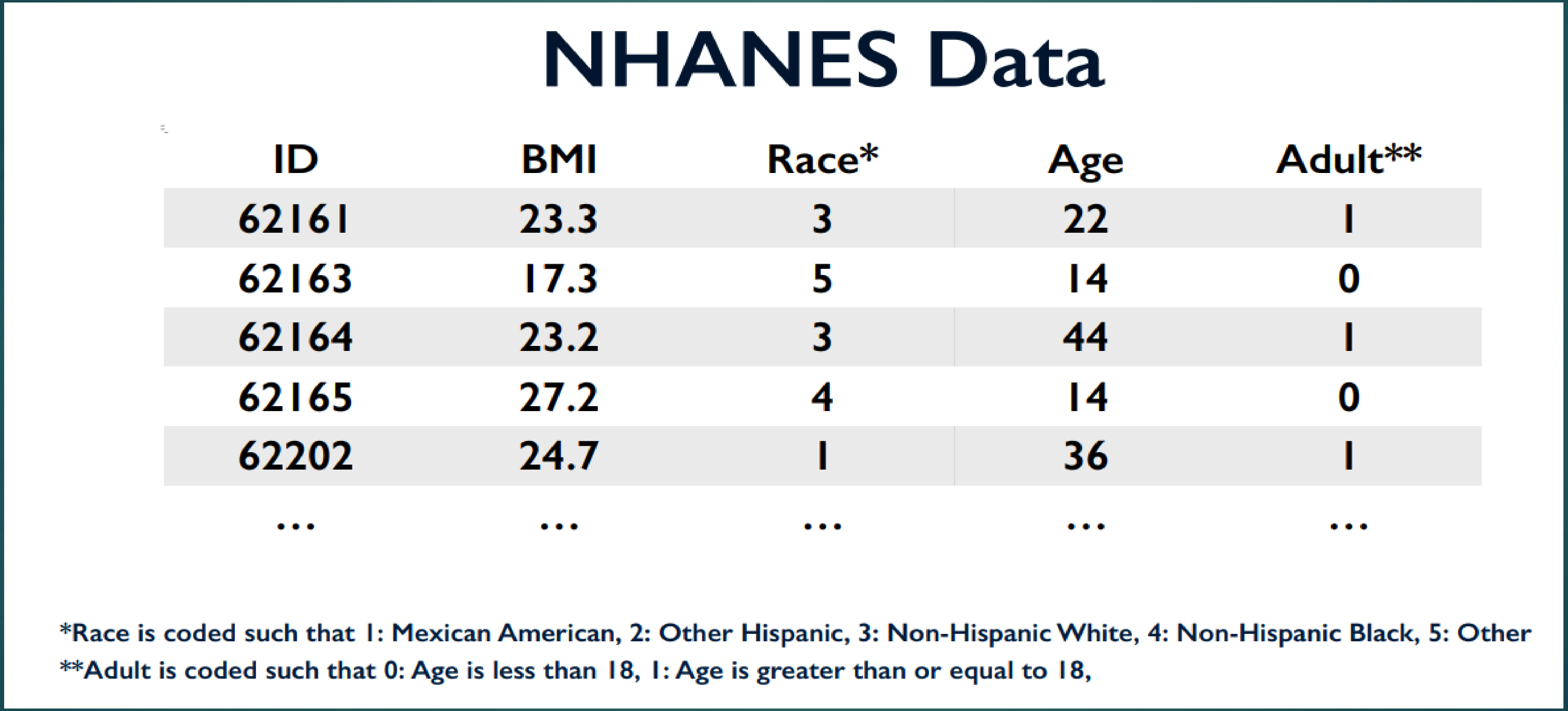
- Categorical (or Qualitative) Variables

Classifies individuals or items into different groups.

- Ordinal – groups have an order or ranking. (Junior, senior).
- Nominal – groups are merely names, no ranking.(

Medical Data

National Health and Nutrition Examination



NHANES Data

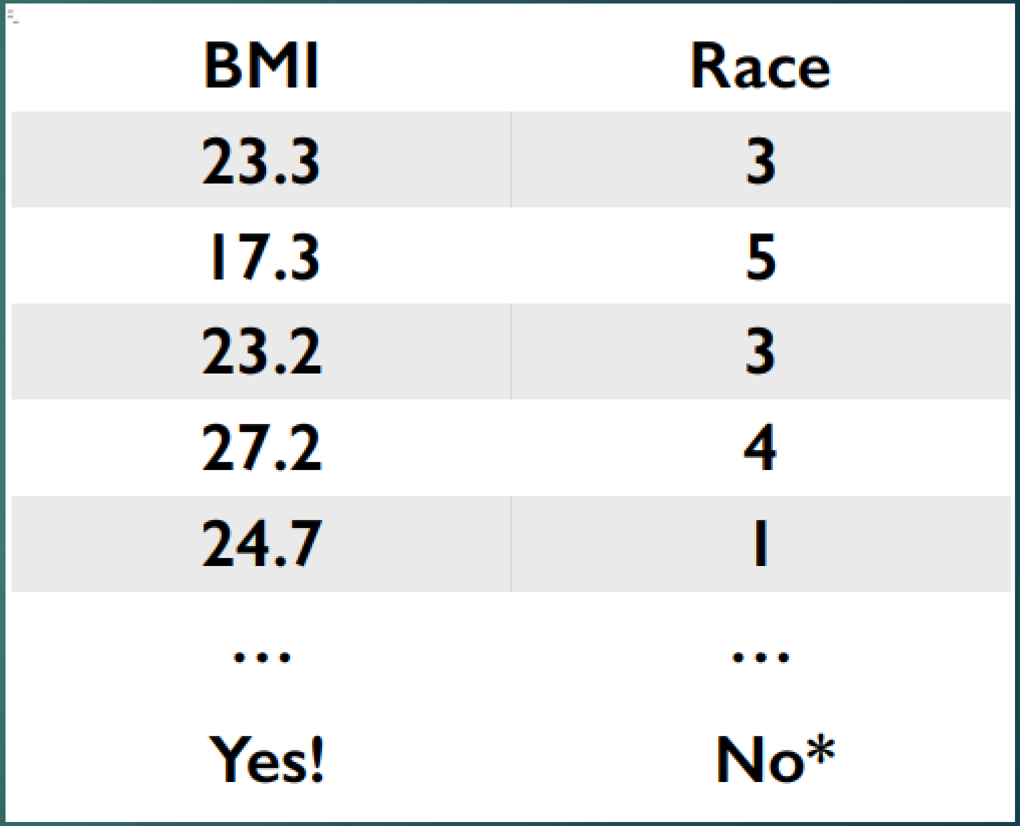
ID	BMI	Race*	Age	Adult**
62161	23.3	3	22	1
62163	17.3	5	14	0
62164	23.2	3	44	1
62165	27.2	4	14	0
62202	24.7	1	36	1
…	…	…	…	…

*Race is coded such that 1: Mexican American, 2: Other Hispanic, 3: Non-Hispanic White, 4: Non-Hispanic Black, 5: Other

**Adult is coded such that 0: Age is less than 18, 1: Age is greater than or equal to 18,

Think about It

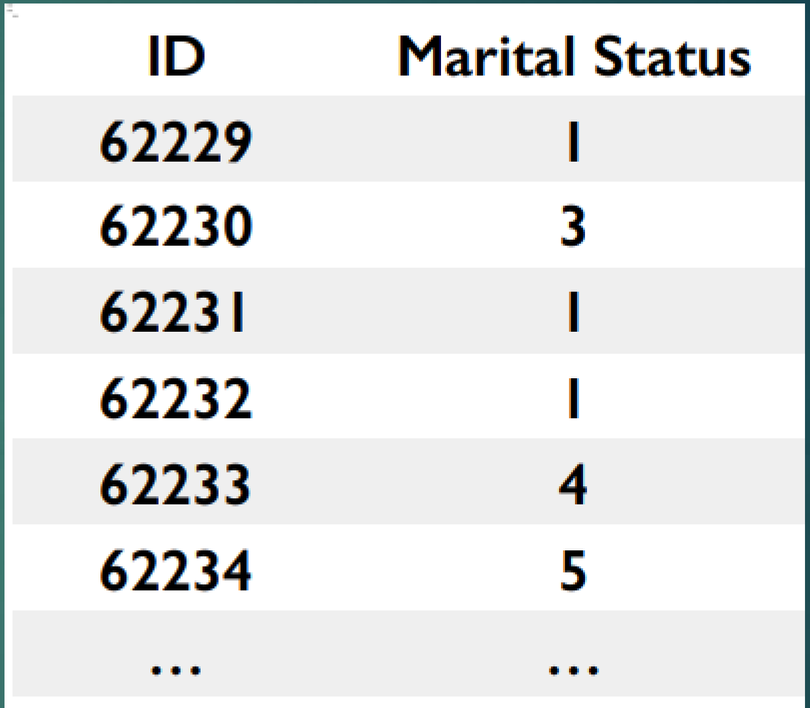
- Could we reasonably compute the average response for each of these two variables?



BMI	Race
23.3	3
17.3	5
23.2	3
27.2	4
24.7	1
…	…
Yes!	No*

Categorical Data

- Classifies individuals or items into different groups

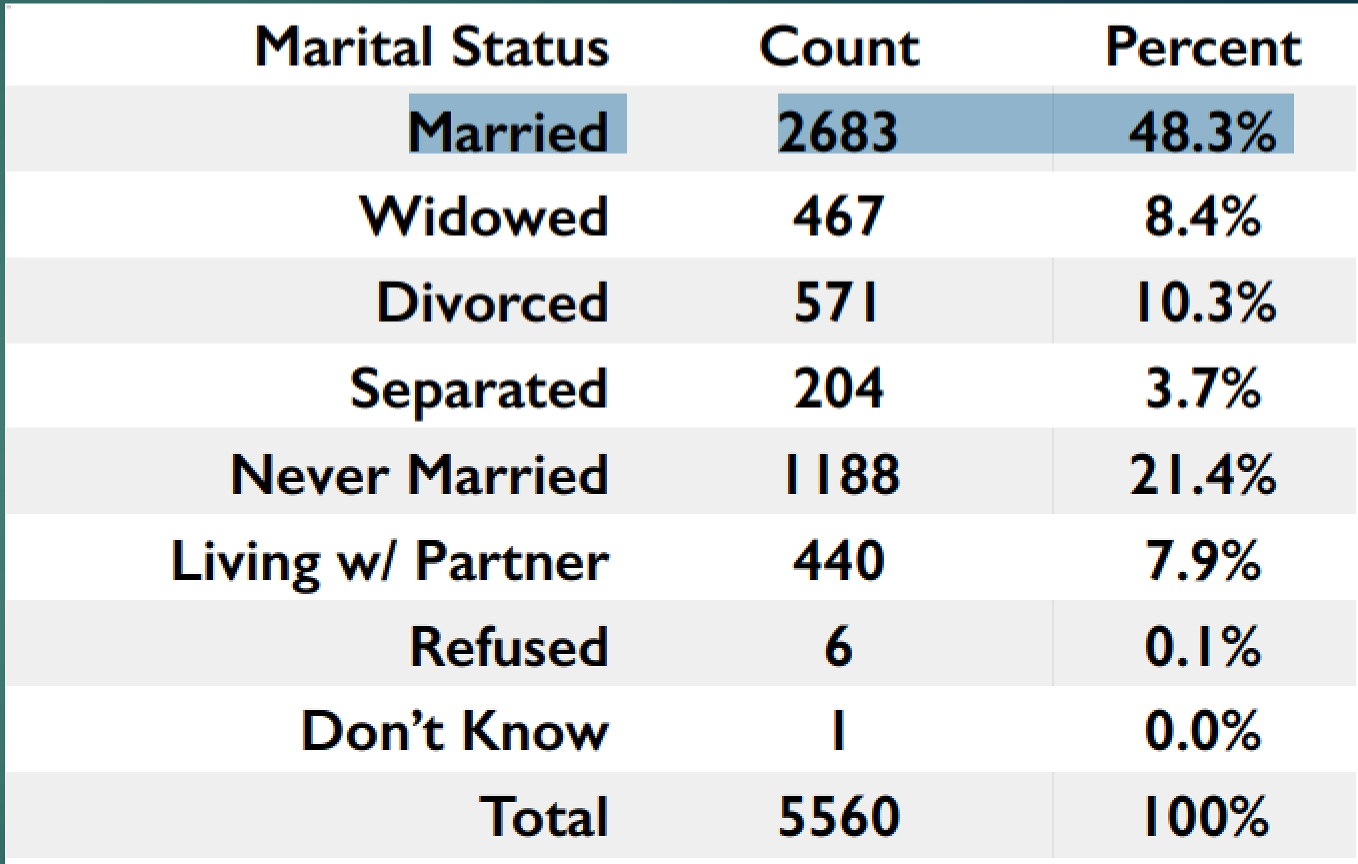


ID	Marital Status
62229	1
62230	3
62231	1
62232	1
62233	4
62234	5
…	…

*Marital Status is coded such that… 1: Married, 2: Widowed, 3: Divorced, 4: Separated, 5: Never Married, 6: Living with Partner, 7: Refused, 8: Don’t Know

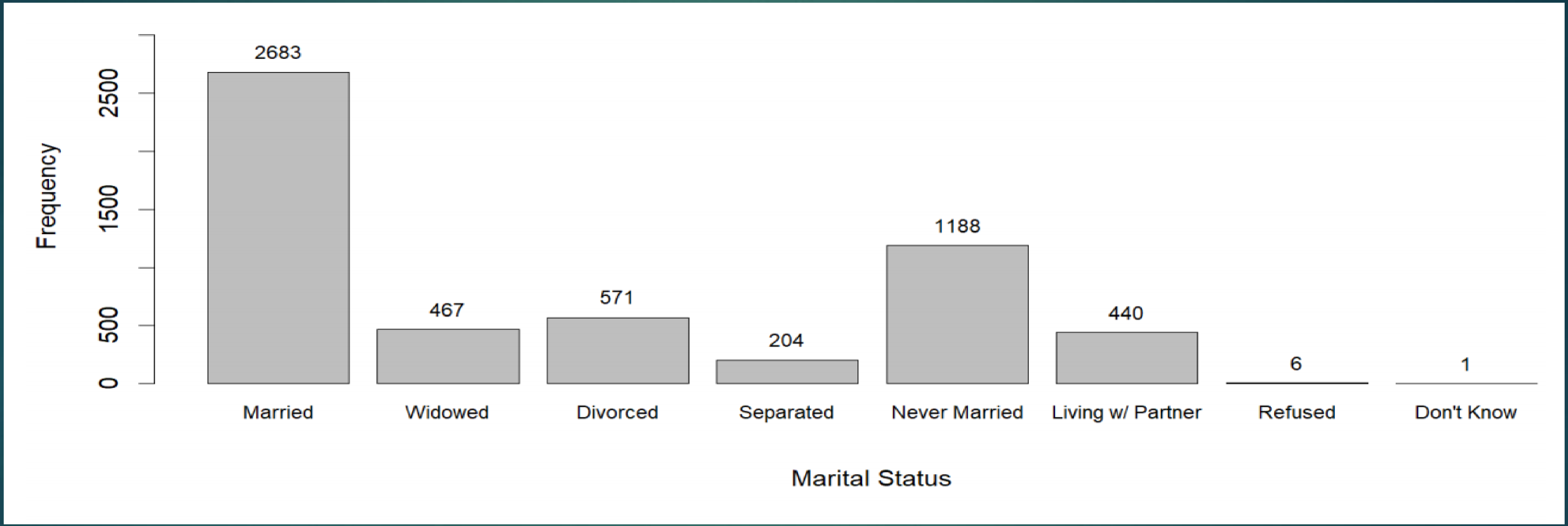
Frequency Table

- Counts
- Percentages

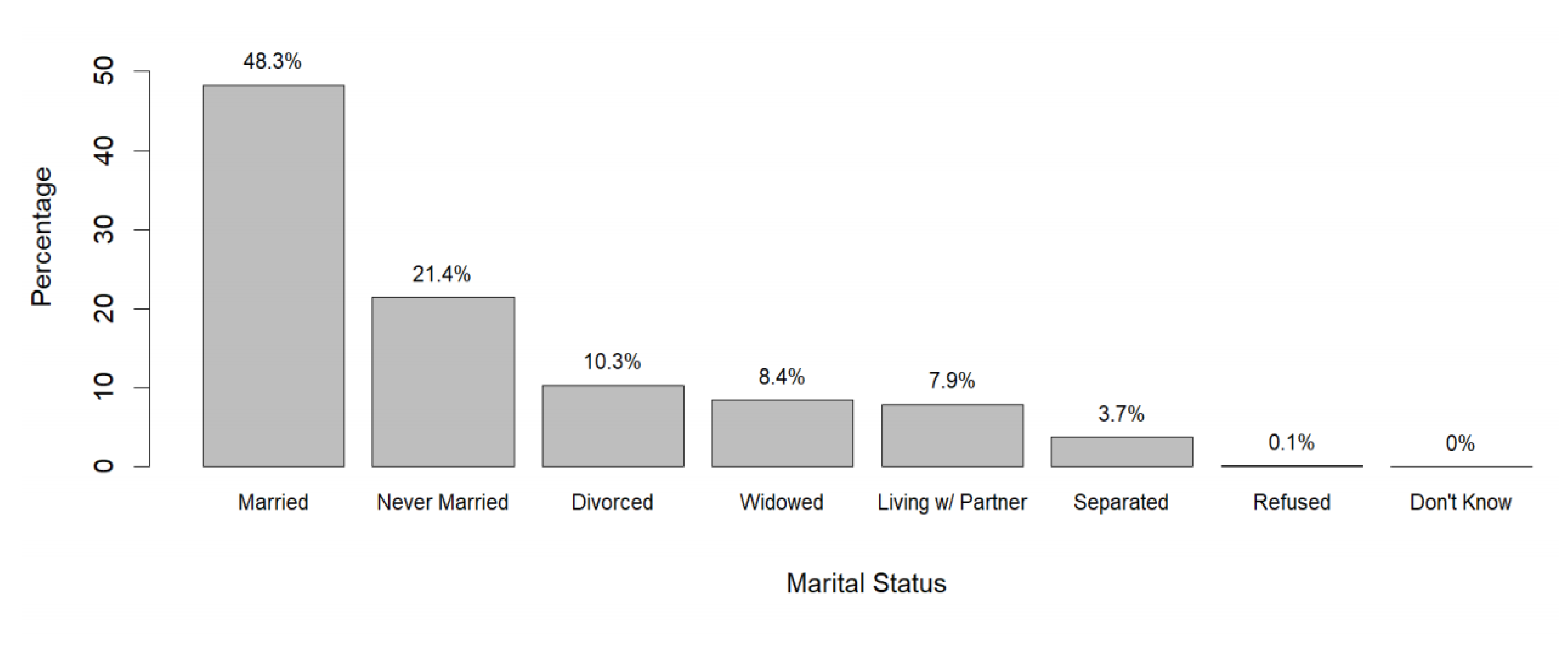


Marital Status	Count	Percent
Married	2683	48.3%
Widowed	467	8.4%
Divorced	571	10.3%
Separated	204	3.7%
Never Married	1188	21.4%
Living w/ Partner	440	7.9%
Refused	6	0.1%
Don’t Know	1	0.0%
Total	5560	100%

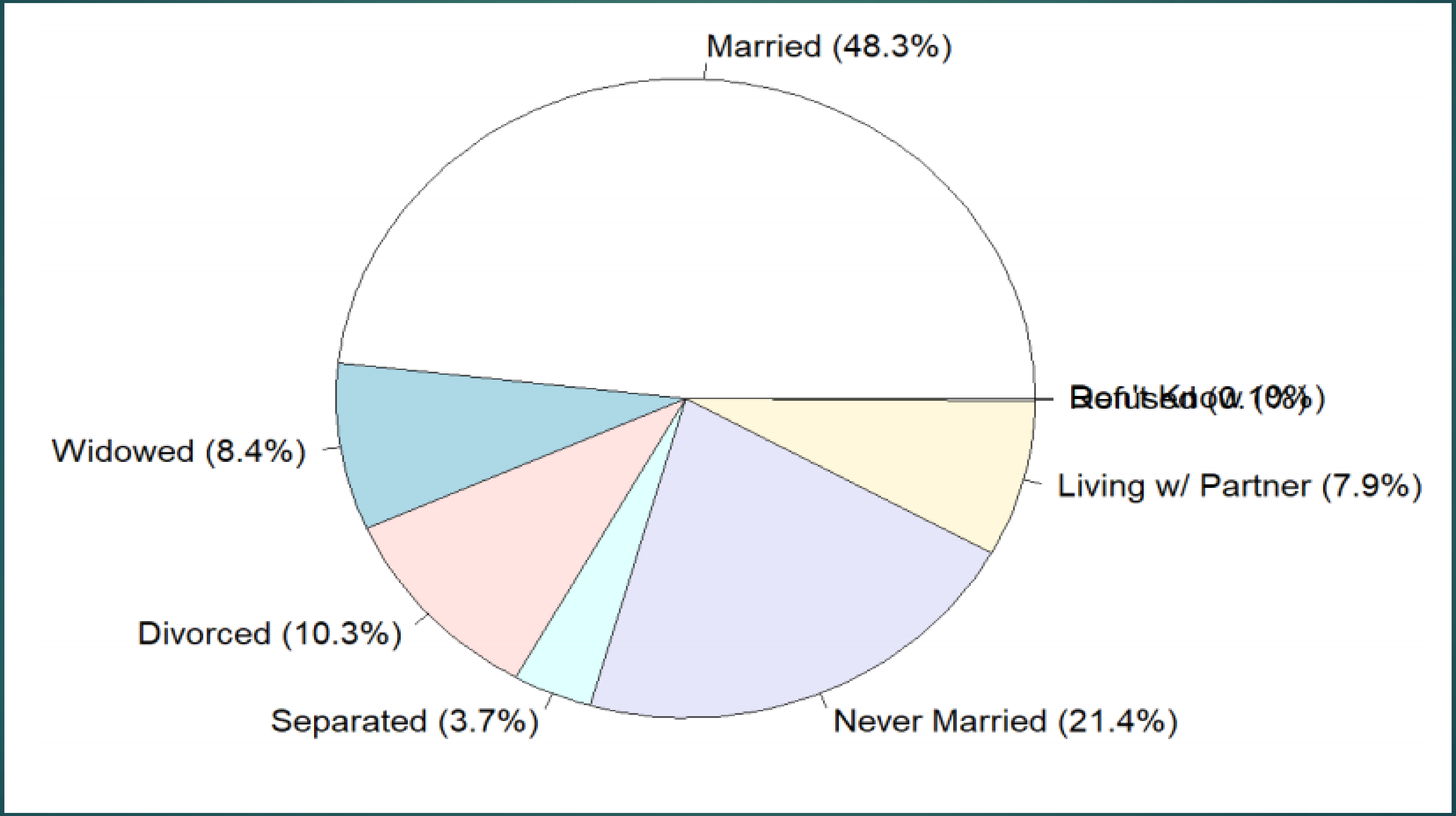
Bar Chart of Marital Status



Bar Chart of Marital Status



Pie Chart of Marital Status



Categorical Data

- Frequency Tables

– Great for numerical summaries

- Bar Charts

– Great for visualization

- Pie Charts

– Use with caution

Quantitative Data: Histograms

- What are Quantitative Variables?

Variables that have a numerical value (quantity) that we can perform mathematical operations on.

Examples: Height, weight, income, test scores, shoe size, number of “heads” after 10 coin flips

Adult Male Heights

Shape

Overall appearance of histogram. Can be symmetric, bell-shaped, left skewed, right skewed, etc.

Center

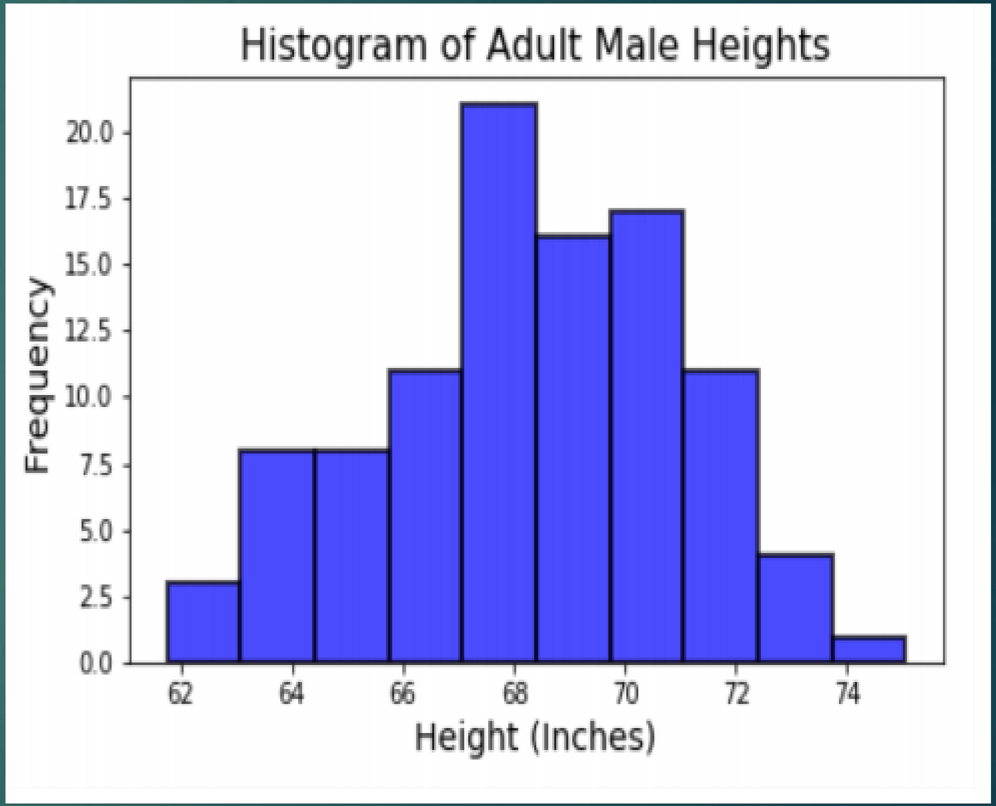
Mean or Median

Spread

How far our data spreads. Range, Interquartile Range (IQR), standard deviation, variance.

Outliers

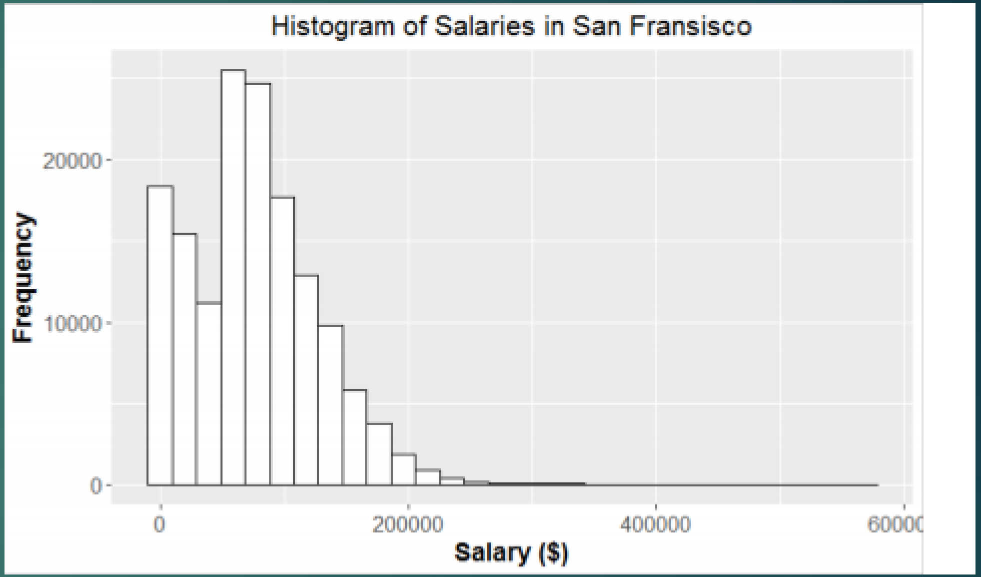
Data points that fall far from the bulk of the data



The distribution of adult male heights is roughly bell shaped with a center of about 68 inches, a range of 13 inches (62 to 75), and no apparent outliers.

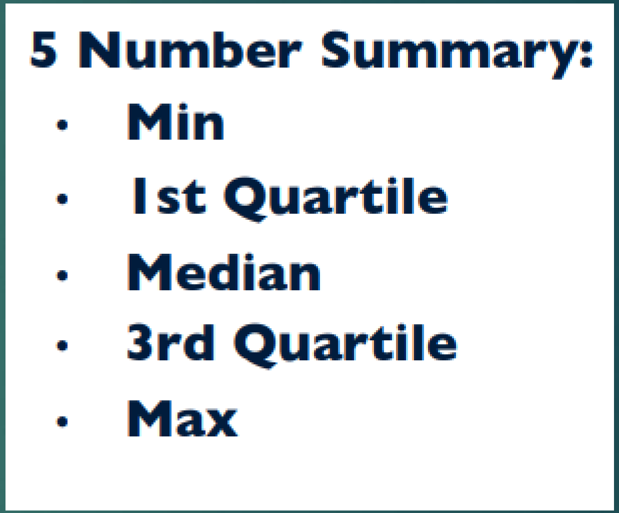
Salaries in San Francisco (2011-2014)

The distribution of salaries in San Francisco is bimodal and skewed to the right, centered at about $80,000 with most of the data between $40,000 and $120,000, a range of roughly $600,000, and outliers are present on the higher end.



Numerical Summaries

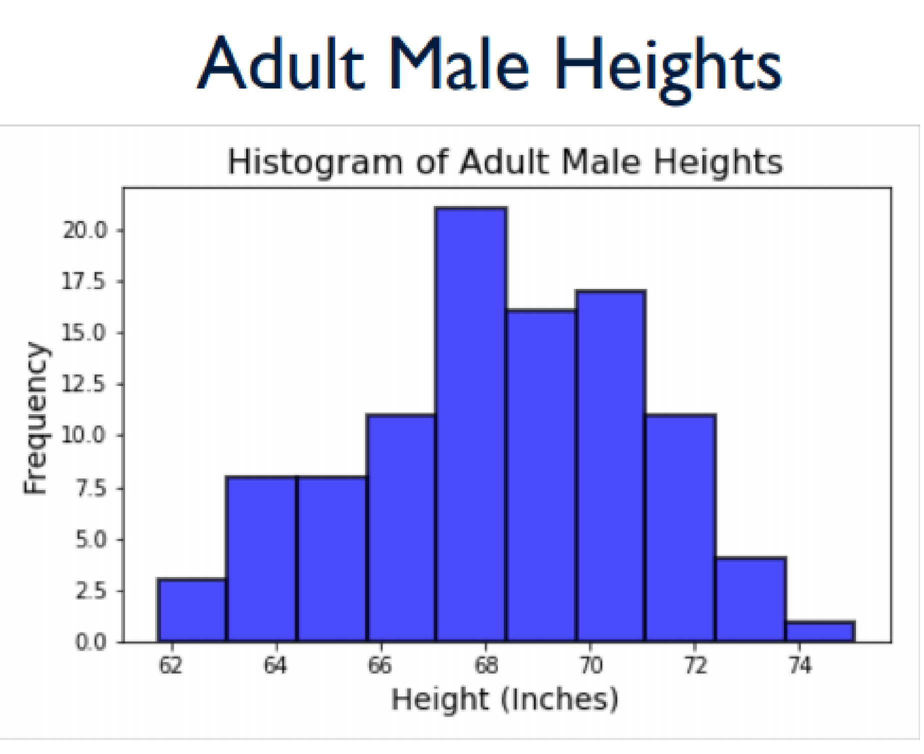
(also called summary statistics) are used alongside our graphical representation of data to give a first impression of what our data looks like.



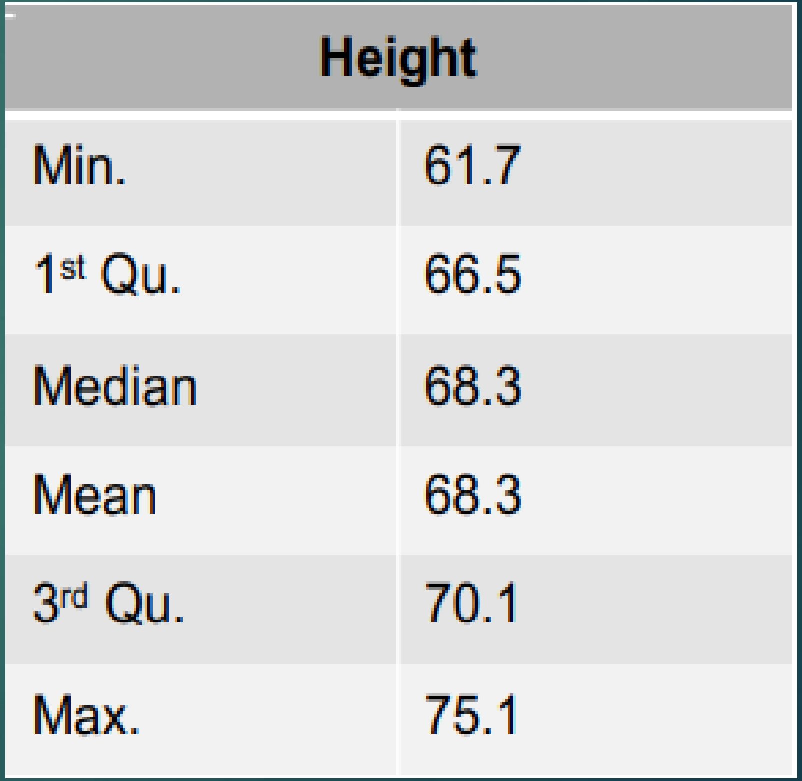
5 Number Summary:

- **Min**
- **1st Quartile**
- **Median**
- **3rd Quartile**
- **Max**

Numerical Summaries

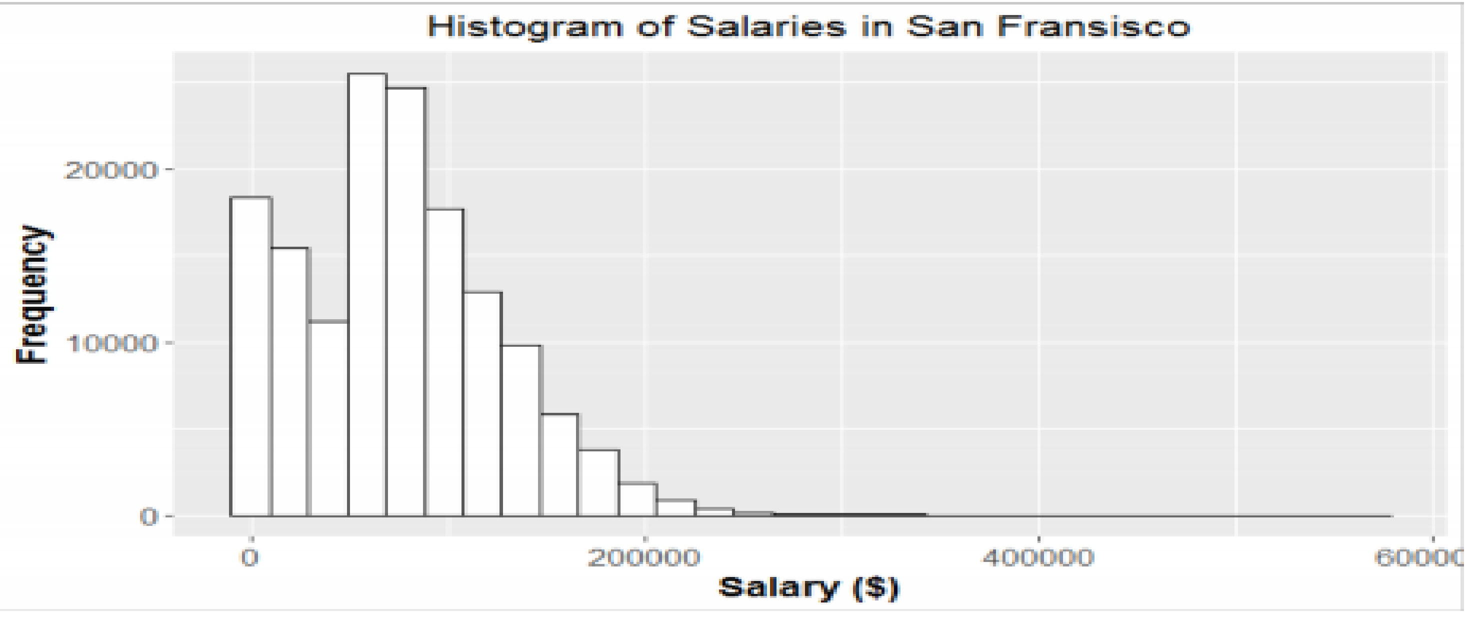


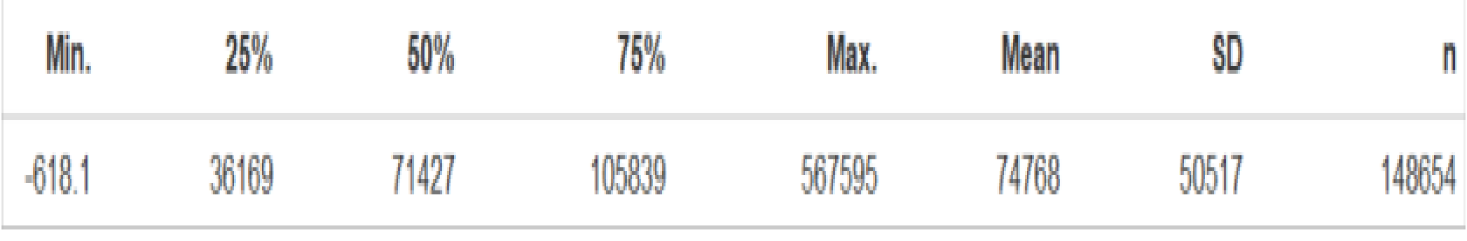
Adult Male Heights



Height	
Min.	61.7
1st Qu.	66.5
Median	68.3
Mean	68.3
3rd Qu.	70.1
Max.	75.1

Salaries in San Francisco (2011-2014)



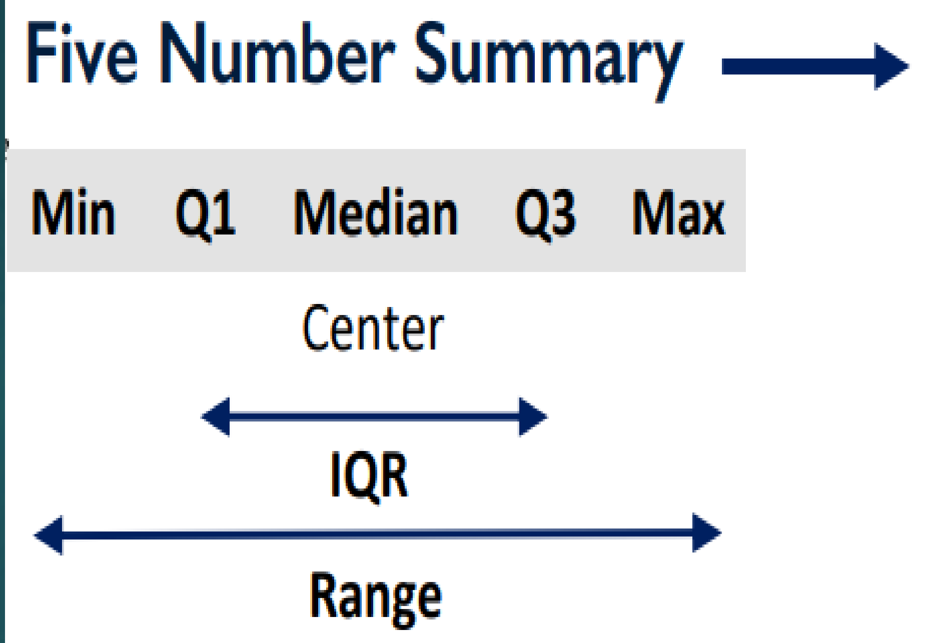


Min.	25%	50%	75%	Max.	Mean	SD	n
-618.1	36169	71427	105839	567595	74768	50517	148654

Quantitative Data Graphical Summary: Boxplots

- Boxplots provide a graphical picture of the five-number summary: showing center (median), spread (IQR and range), and identifies potential outliers.
- Boxplots can hide some shape aspects (histograms do better job at displaying shape)
- Side-by-Side Boxplots are useful for comparing two or more sets of observations.

What is a Boxplot?



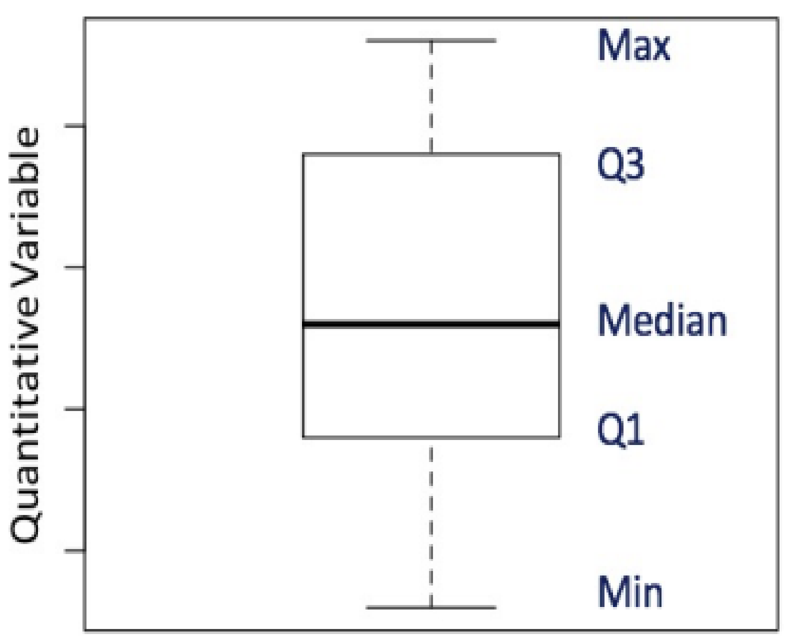
Five Number Summary

Min Q1 Median Q3 Max

Center

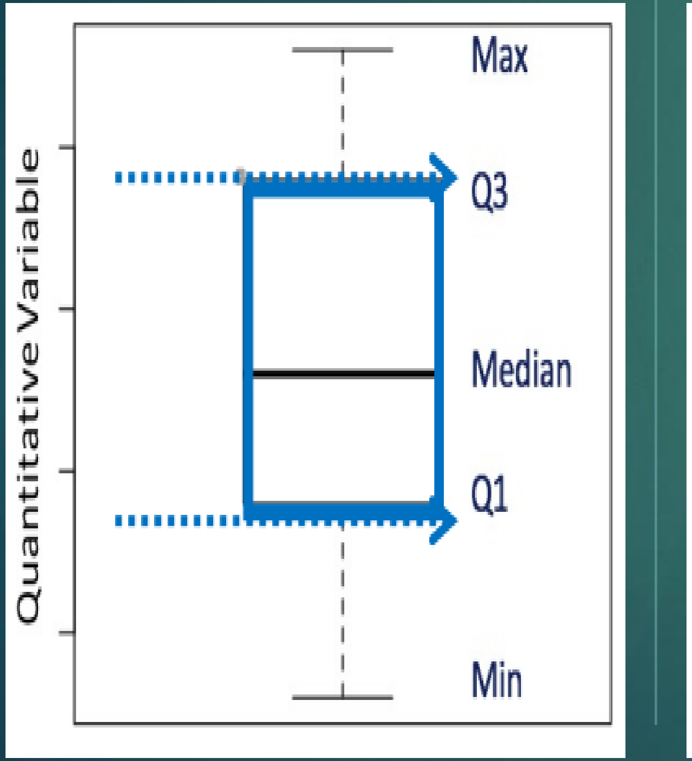
IQR

Range

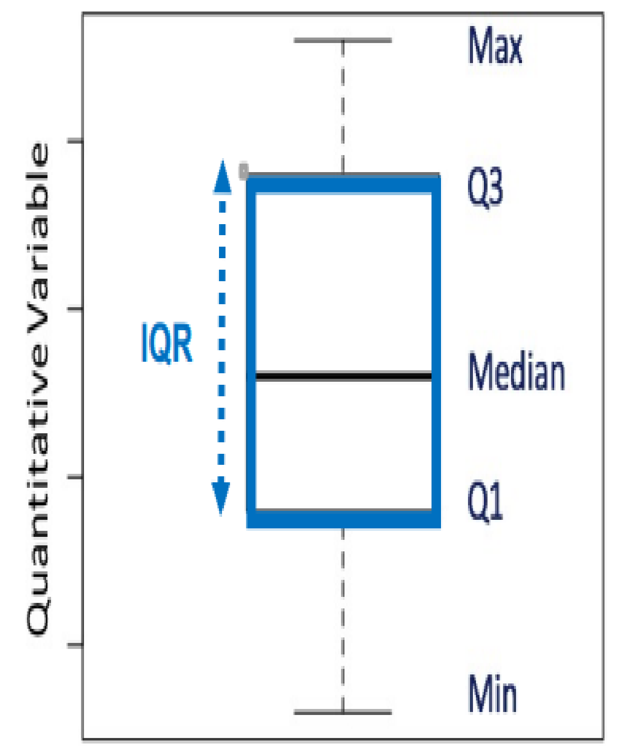


What is a Boxplot?

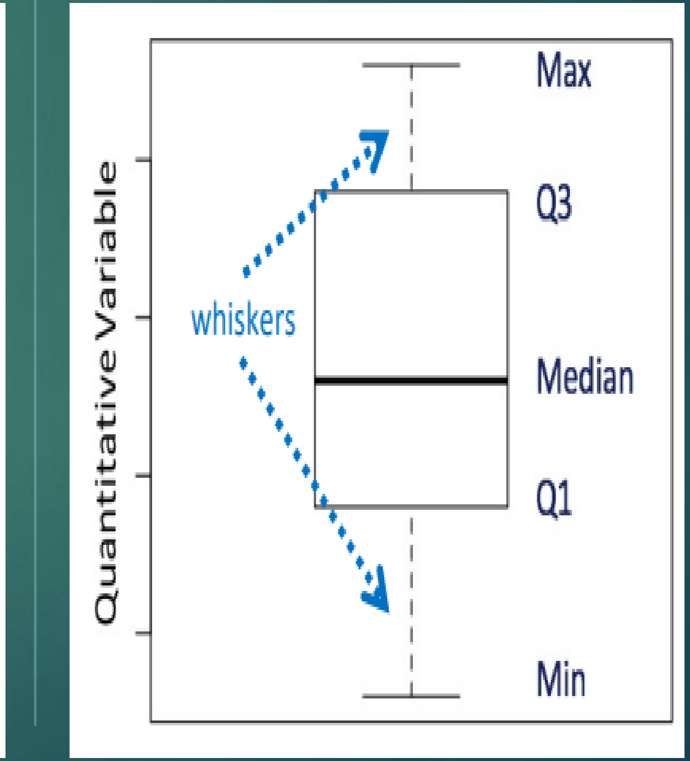
1St & 3rd Quartile



Inter Quartile Range



Whisker



Visualization with Matplotlib

- We'll now take an in-depth look at the Matplotlib tool for visualization in Python. Matplotlib is a multiplatform data visualization library built on NumPy arrays, and designed to work with the broader SciPy stack. It was conceived by John Hunter in 2002, originally as a patch to IPython for enabling interactive MATLAB-style plotting via gnu plot from the IPython command line.

Importing matplotlib

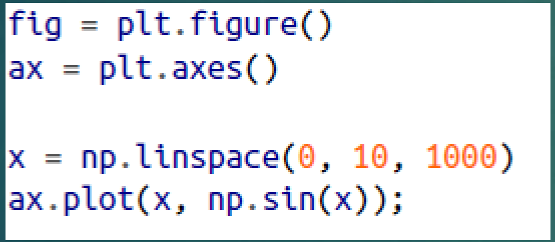
- Just as we use the np shorthand for NumPy and the pd shorthand for Pandas, we will use some standard shorthands for Matplotlib imports:

```
import matplotlib as mpl
import matplotlib.pyplot as plt
```

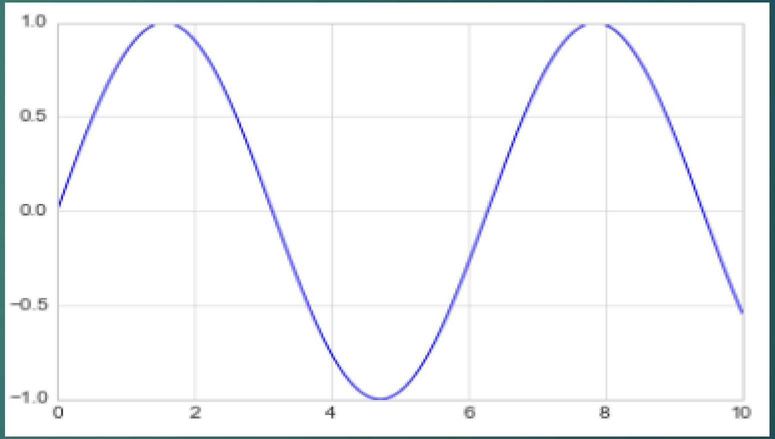
Setting Styles

We will use the plt.style directive to choose appropriate aesthetic styles for our figures.
Here we will set the classic style, which ensures that the plots we create use the
classic Matplotlib style:

plt.style.use('classic')

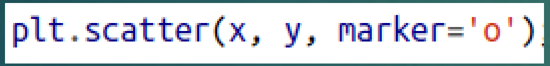


```
fig = plt.figure()
ax = plt.axes()

x = np.linspace(0, 10, 1000)
ax.plot(x, np.sin(x));
```

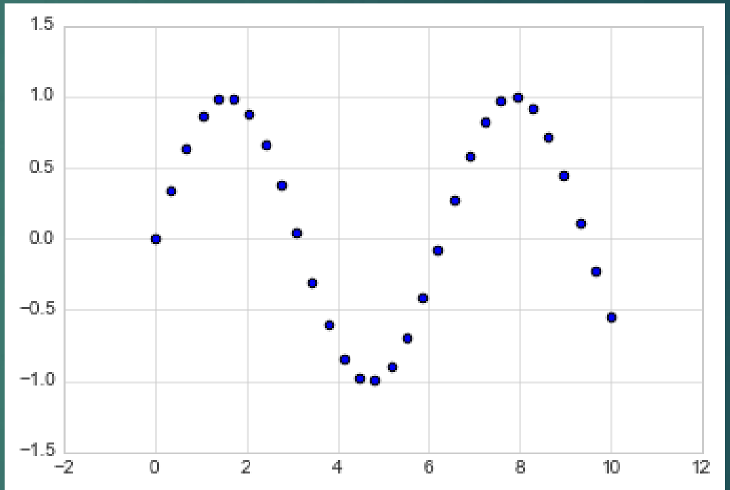


Scatter Plots with plt.scatter

A second, more powerful method of creating scatter plots is the plt.scatter
function,
which can be used very similarly to the plt.plot function

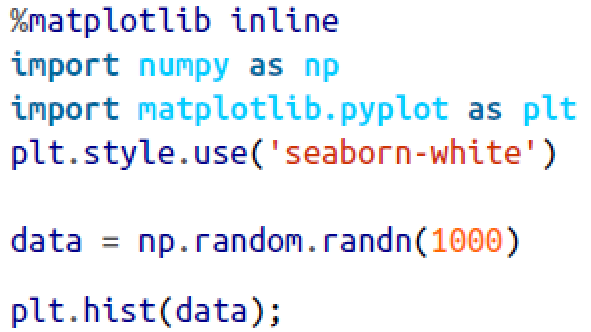


```
plt.scatter(x, y, marker='o')
```

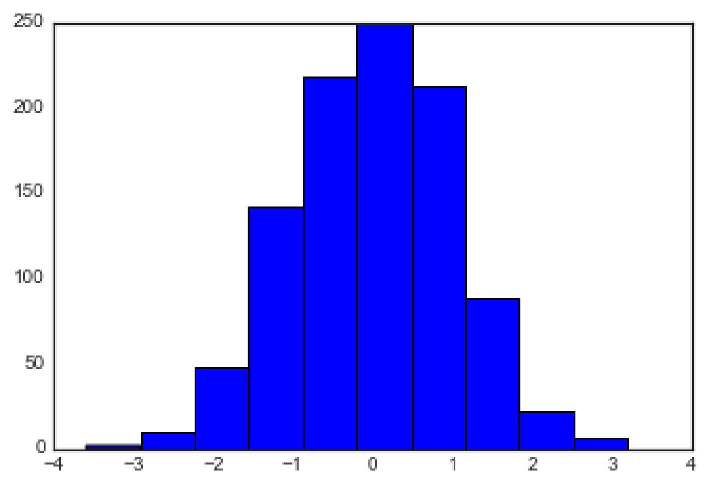


Histograms

A simple histogram can be a great first step in understanding a dataset.

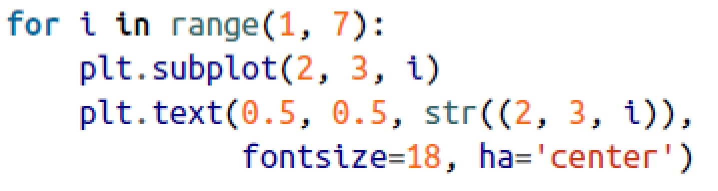


```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
plt.style.use('seaborn-white')

data = np.random.randn(1000)

plt.hist(data);
```

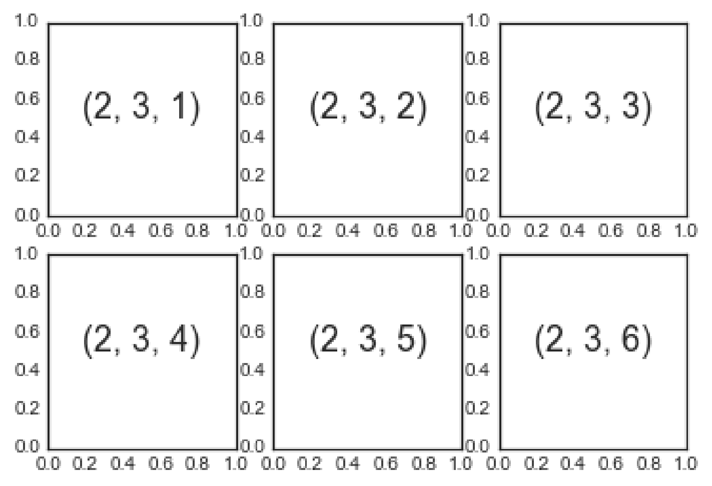


plt.subplot: Simple Grids of Subplots

Aligned columns or rows of subplots are a common enough need that Matplotlib has several convenience routines that make them easy to create. The lowest level of these is plt.subplot(), which creates a single subplot within a grid. As you can see, this command takes three integer arguments—the number of rows, the number of columns, and the index of the plot to be created in this scheme, which runs from the upper left to the bottom right



```
for i in range(1, 7):
    plt.subplot(2, 3, i)
    plt.text(0.5, 0.5, str((2, 3, i)),
             fontsize=18, ha='center')
```



Visualization with Seaborn

Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures. Seaborn aims to make visualization a central part of exploring and understanding data.

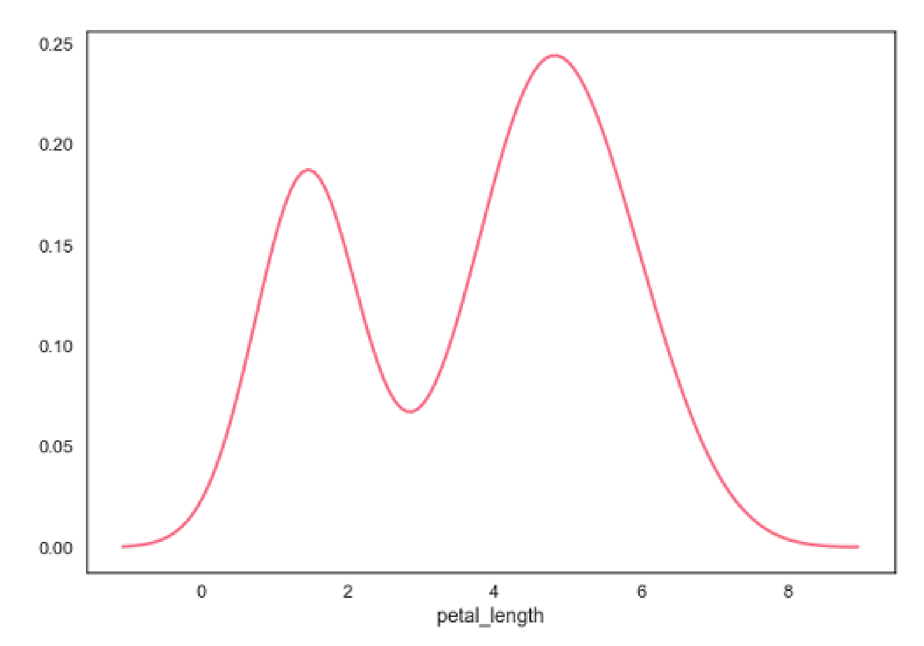
- We import seaborn, which is the only library necessary for this simple example.

```
Import seaborn as sns
```

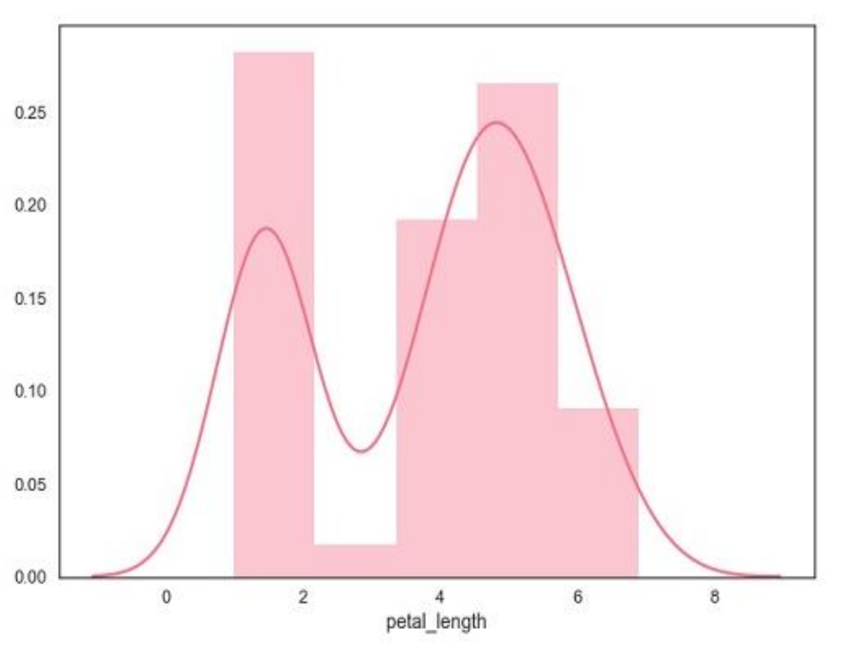
Seaborn - Kernel Density Estimates

Kernel Density Estimation (KDE) is a way to estimate the probability density function of a continuous random variable. It is used for non-parametric analysis

Kernel Density Estimation



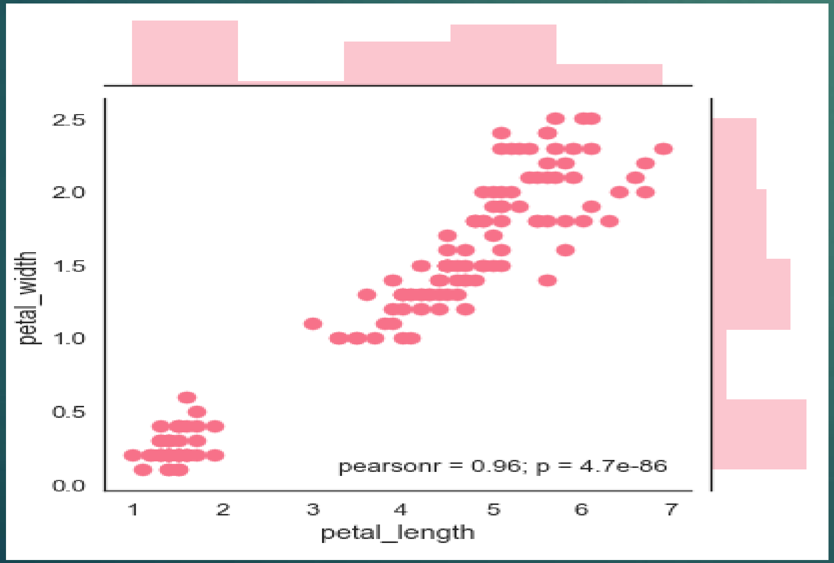
Fitting Parametric Distribution



Visualization with Seaborn

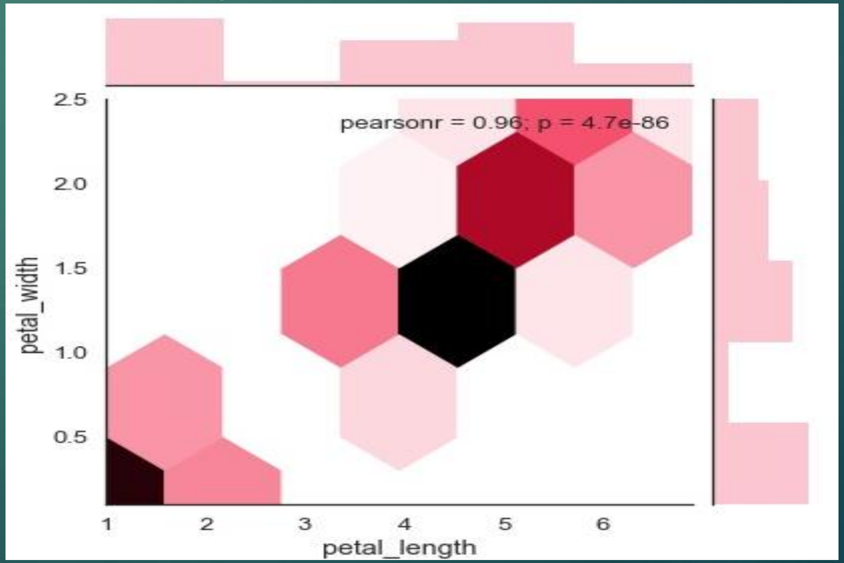
Scatter Plot

- Scatter plot is the most convenient way to visualize the distribution where each observation is represented in two-dimensional plot via x and y axis.



Hexbin Plot

- Hexagonal binning is used in bivariate data analysis when the data is sparse in density i.e., when the data is very scattered and difficult to analyze through scatterplots.



Visualizing Pairwise Relationship

To plot multiple pairwise bivariate distributions in a dataset, you can use the **pairplot()** function. This shows the relationship for (n,2) combination of variable in a DataFrame as a matrix of plots and the diagonal plots are the univariate plots.

Parameter & Description

Data

Dataframe

hue

Variable in data to map plot aspects to different colors.

palette

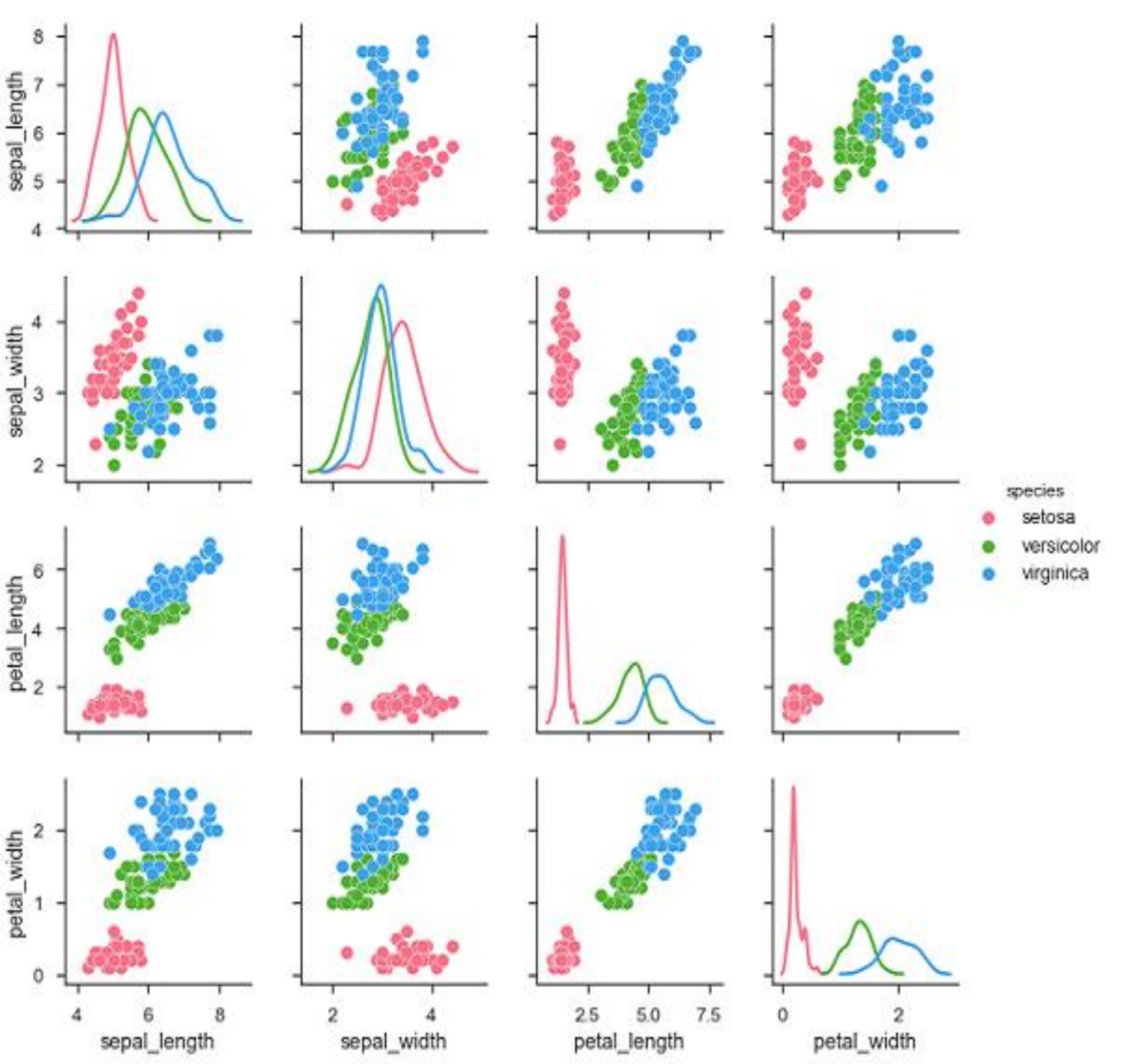
Set of colors for mapping the hue variable.

kind

Kind of plot for the non-identity relationships. {'scatter', 'reg'}

diag_kind

Kind of plot for the diagonal subplots. {'hist', 'kde'}

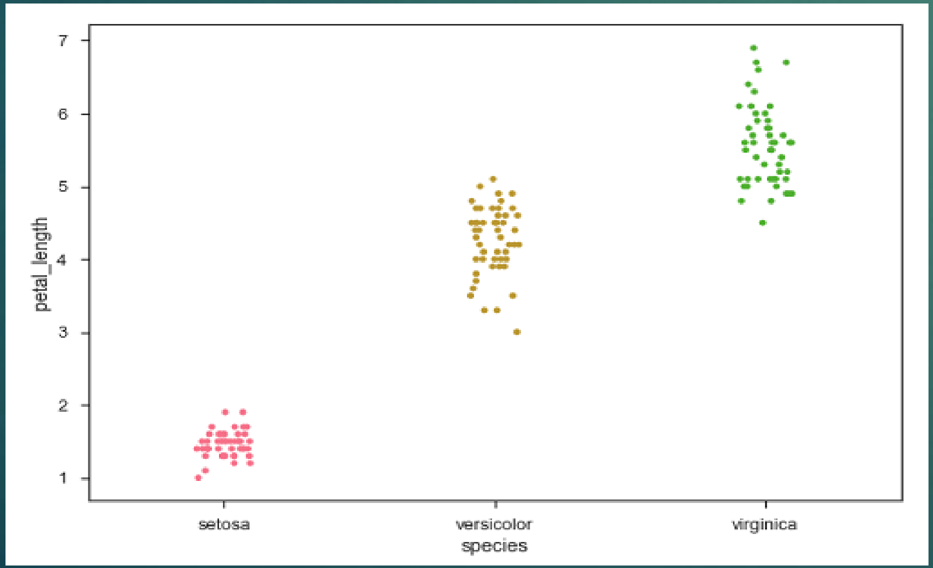


Plotting Categorical Data

When one or both the variables under study are categorical, we use plots like striplot(), swarmplot(), etc,. Seaborn provides interface to do so.

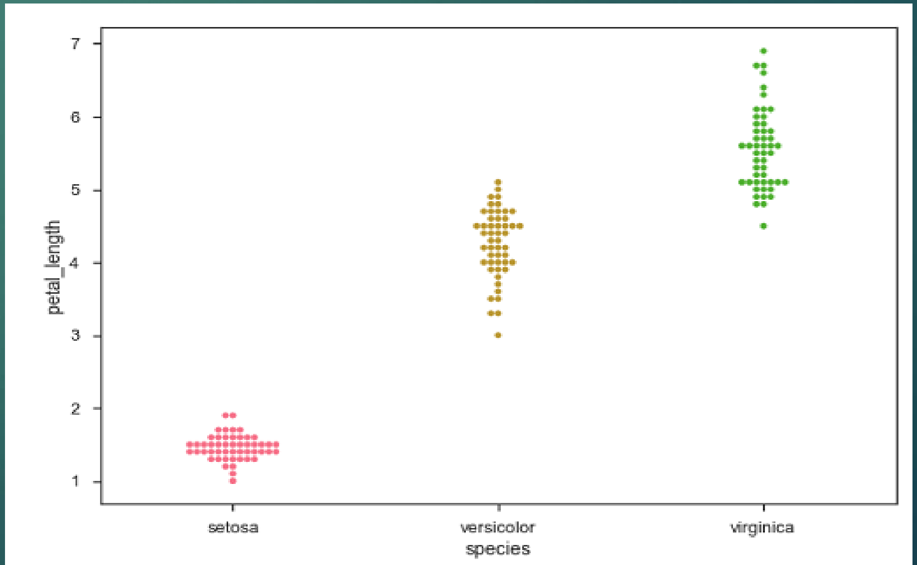
stripplot()

- stripplot() is used when one of the variable under study is categorical. It represents the data in sorted order along any one of the axis.



Swarmplot()

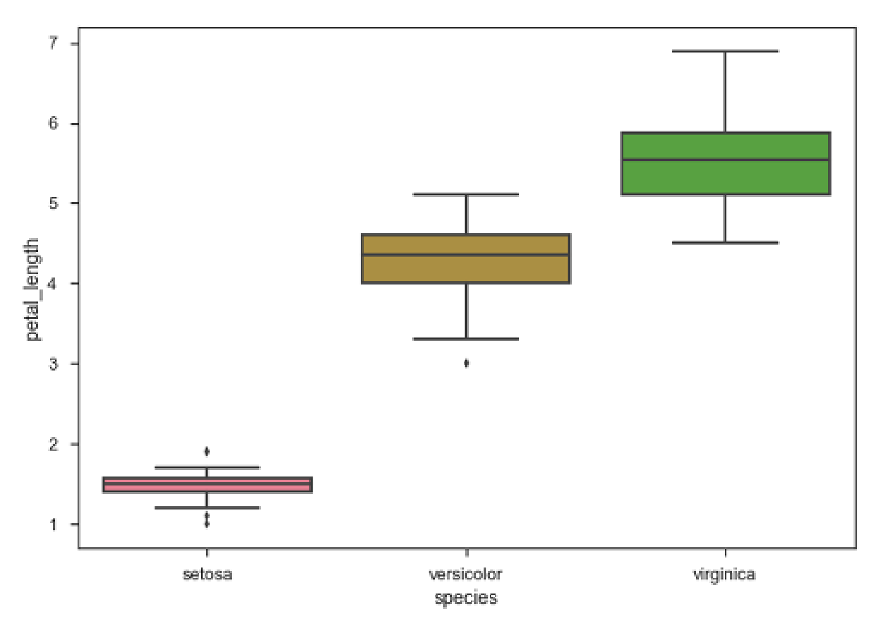
- This function positions each point of scatter plot on the categorical axis and thereby avoids overlapping points.



Distribution of Observations

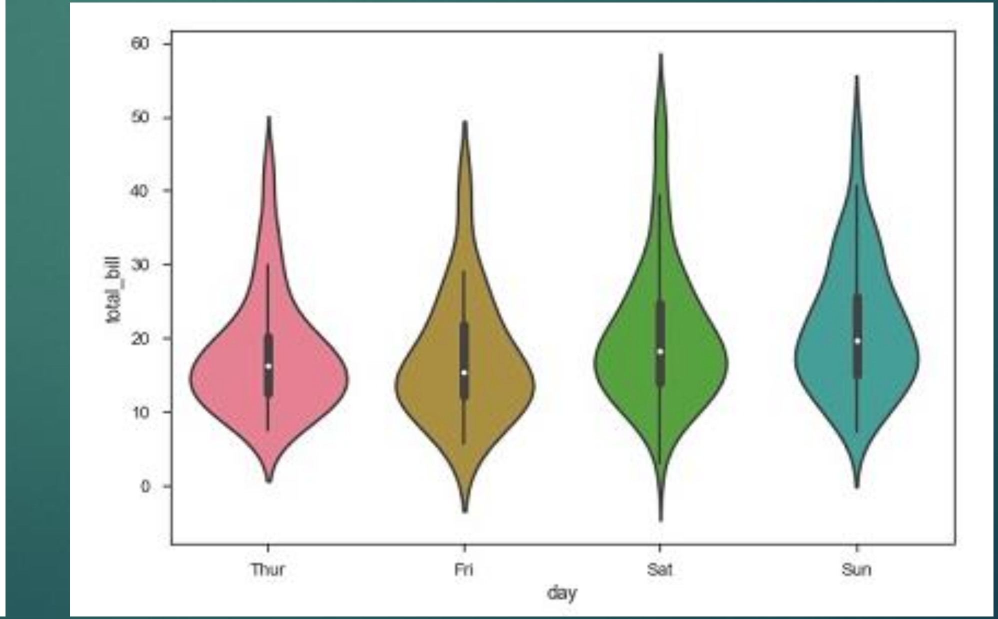
Box Plots

- **Boxplot** is a convenient way to visualize the distribution of data through their quartiles



Violin Plots

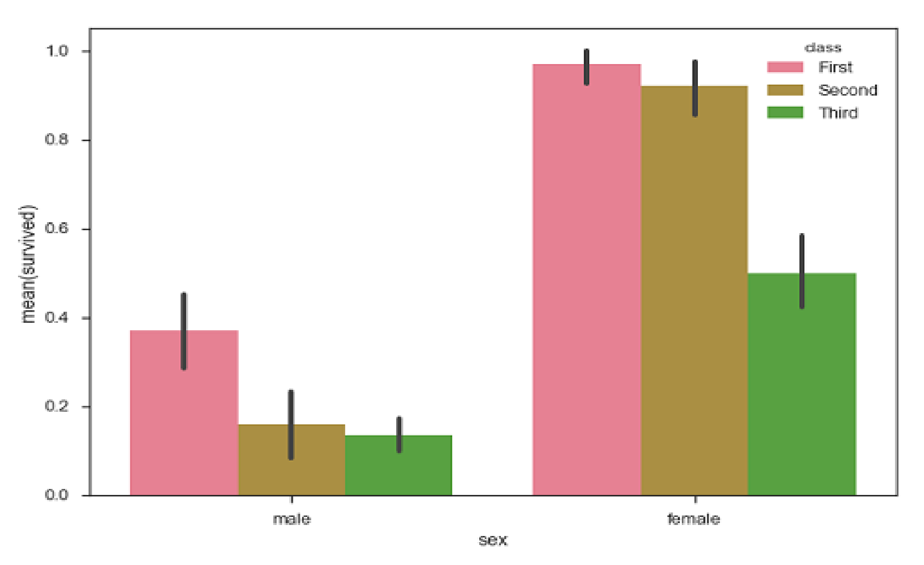
- Violin Plots are a combination of the box plot with the kernel density estimates.



Statistical Estimation

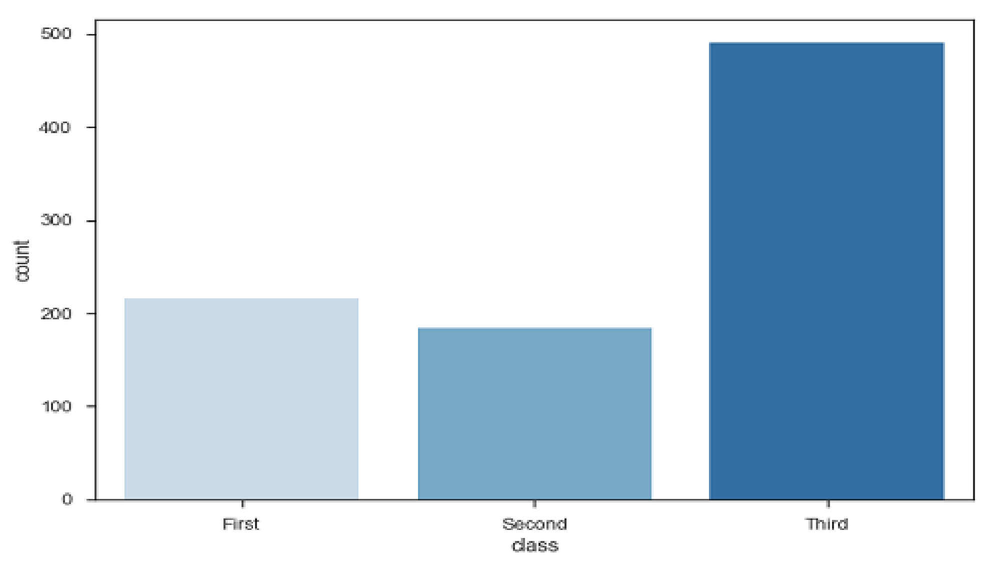
Bar Plot

- The **barplot()** shows the relation between a categorical variable and a continuous variable



countplot()

- A special case in barplot is to show the no of observations in each category rather than computing a statistic for a second variable.



Linear Relationships

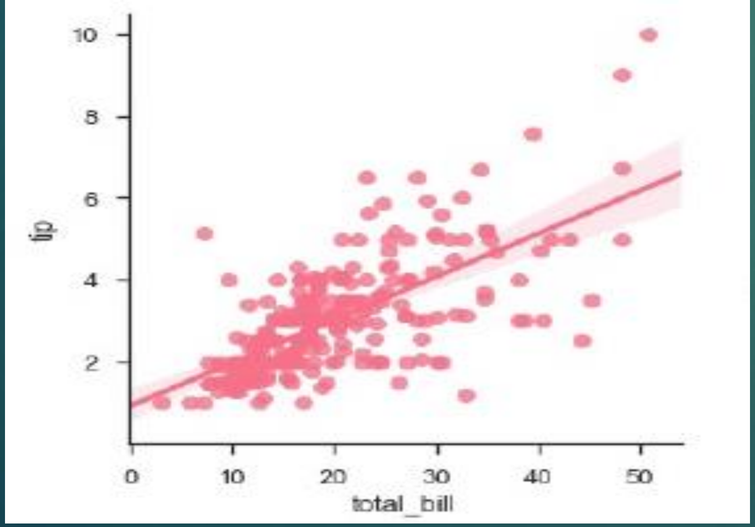
There are two main functions in Seaborn to visualize a linear relationship determined through regression. These functions are **regplot()** and **Implot()**

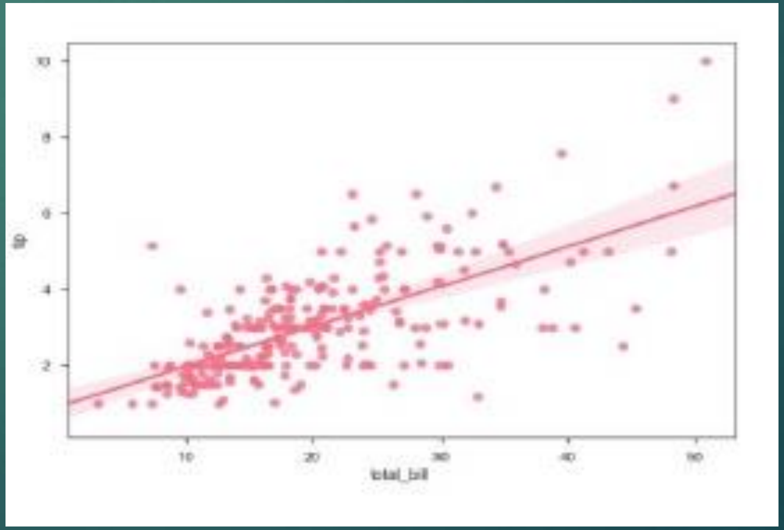
regplot()

accepts the x and y variables in a variety of formats including simple numpy arrays, pandas Series objects

Implot()

data as a required parameter and the x and y variables must be specified as strings.





Thanks