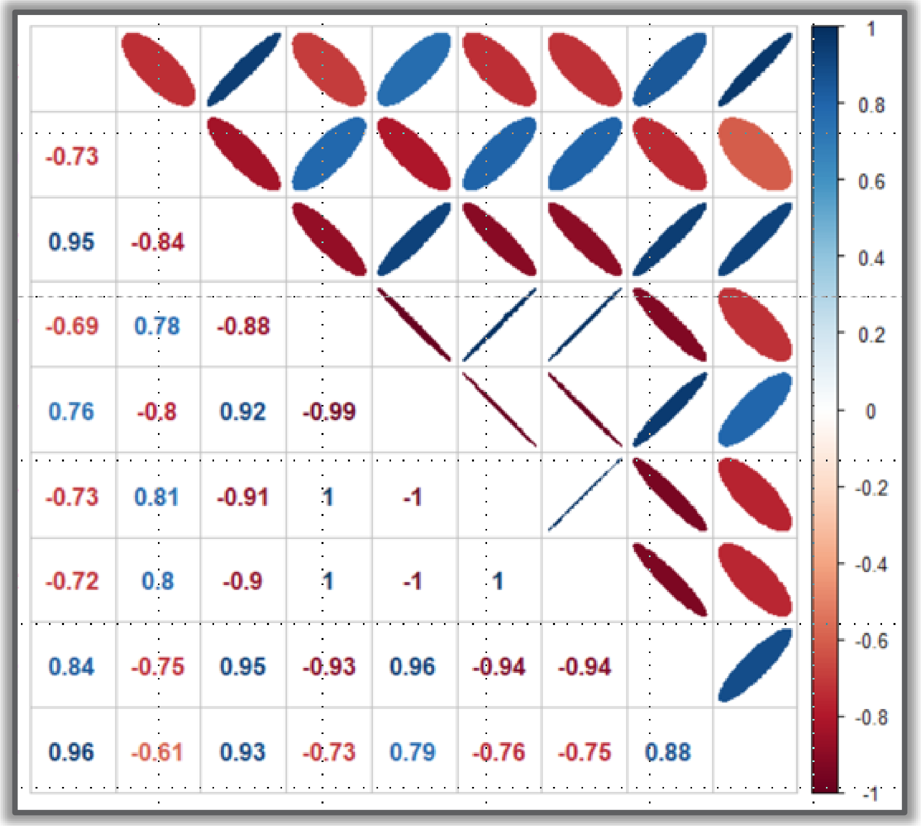
# Bivariate EDA: Correlation & Covariance




- Visualizing pairwise relationships (scatterplot matrix)
- Visualization: Hexagonal Binning and Contours
- Centering and standardizing
- Correlation and covariance matrix and visualization
- Statistical significance of the correlation

**Prof. Nagiza F. Samatova**
samatova@csc.ncsu.edu
Department of Computer Science
North Carolina State University

NC STATE Executive Education

February 2019

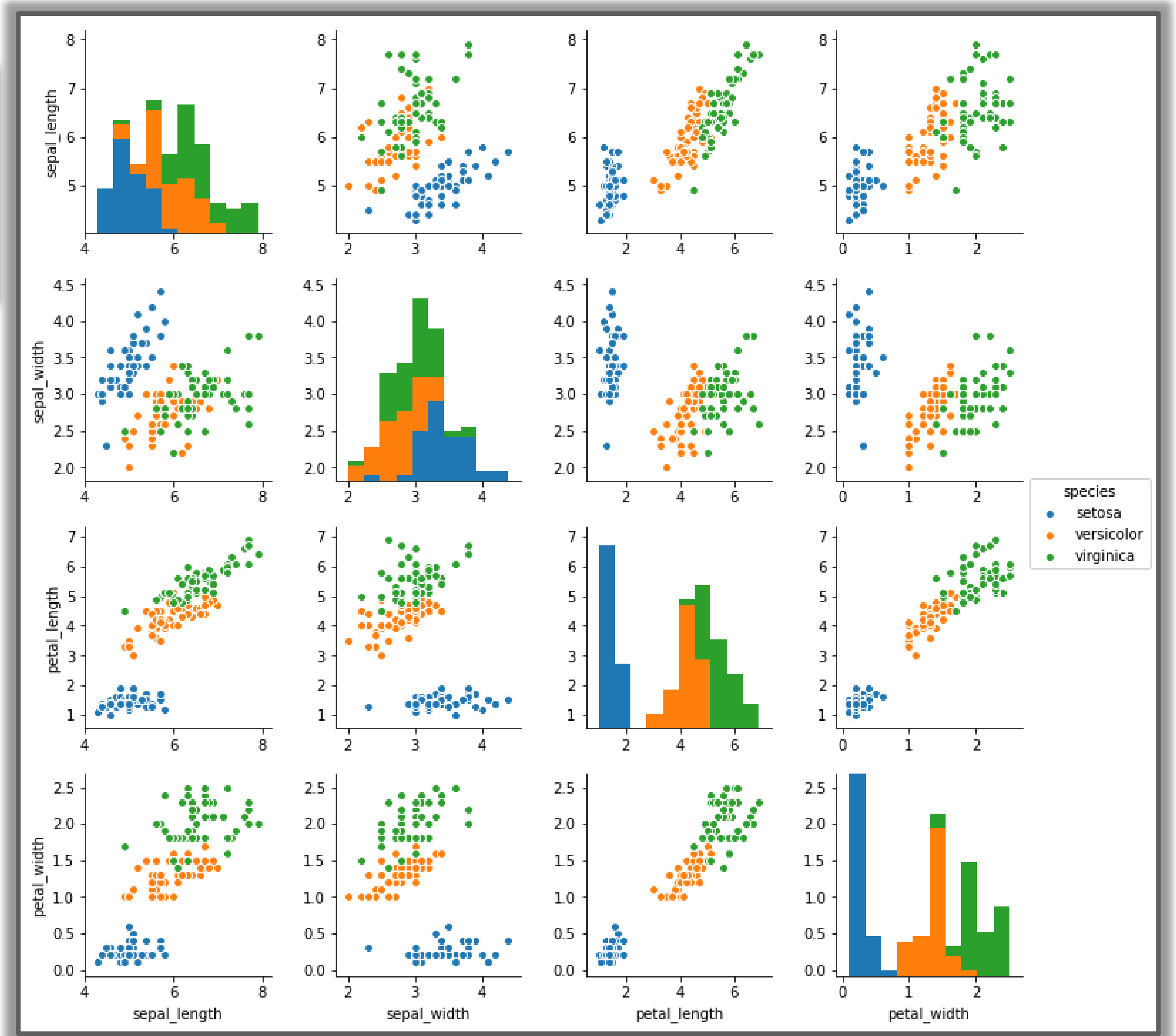
# Bivariate Analysis: Correlation, Covariance, Scatter

| Measure | Description | Comments |
|---|---|---|
| Correlation coefficient | metric that measures the extent to which two numeric variables are associated with one another | ranges from -1 to +1<br>cor (x,y) |
| Correlation matrix | table where variables are shown on both rows and columns; cell values are correlations between variables | corrplot::corrplot(), corrplot::corrplot.mixed(), |
| Correlation test | test that measures statistical significance of the correlation coefficient | cor.test() |
| Scatterplot | plot in which the x-axis is the value of one variable, and the y-axis is the value of the other variable | car::scatterplotMatrix(), gpairs::gpairs() |
| Centering | subtracting the mean from original values | xc = x – mean(x) |
| Covariance | average association between two centered variables | cov (x,y) |
| Correlation | covariance scaled by sd(x) * sd(y); Pearson correlation | cor (x,y) = cov(x,y) / (sd(x)*sd(y)) |
| Z-score, z | centered variable divided by its sd(x): z = xc /sd(x) | mean (z) = 0, sd(z) = 1 |
| Hexagonal binning | plot of two numeric variables with the records binned into hexagons | |
| Contour plot | plot showing the density of two numeric variables | like a topographic map |

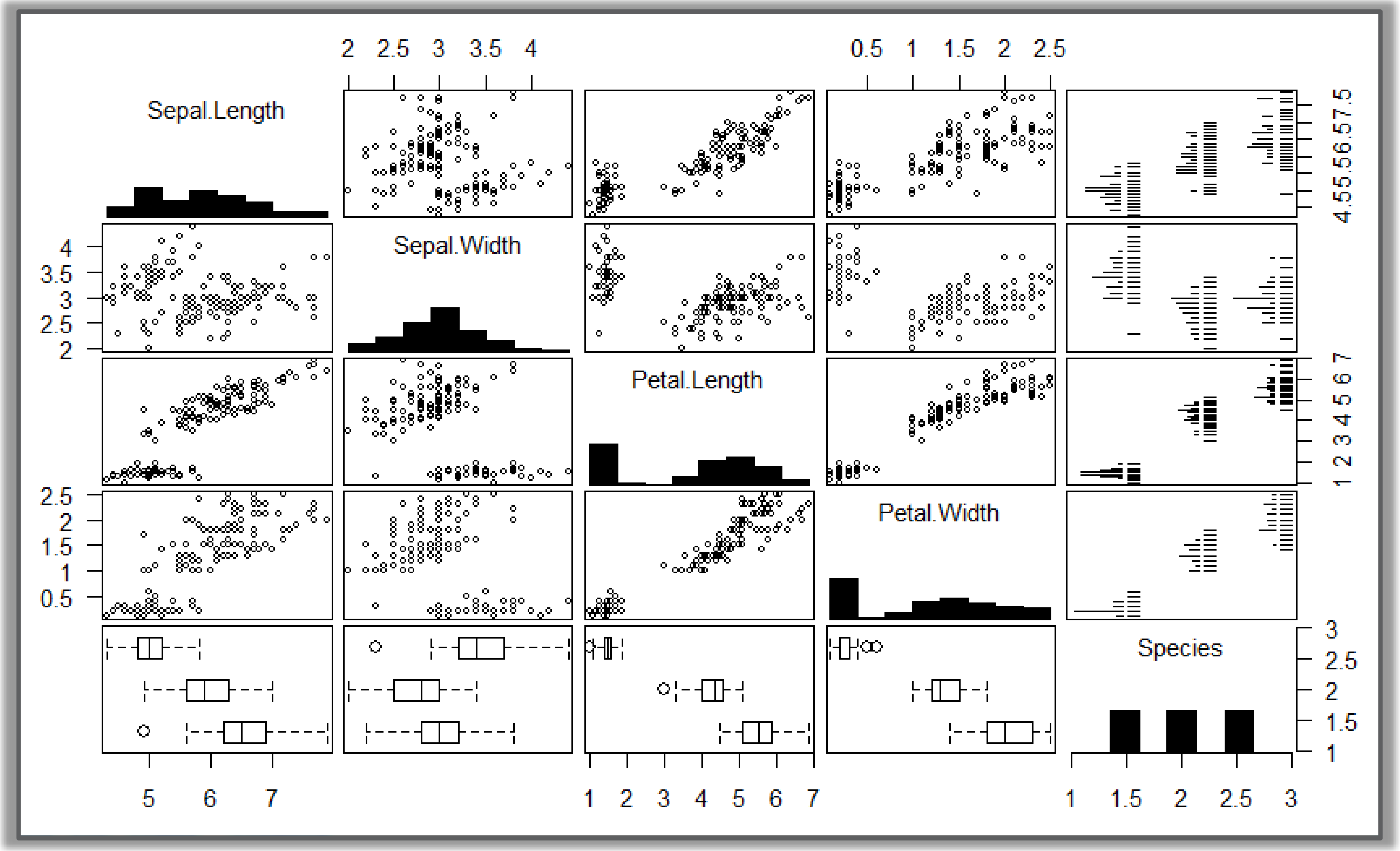
# Visualizing Pairwise Relationships

## Scatterplot Matrix

```
import seaborn as sns
df = sns.load_dataset("iris")
sns.pairplot(df, hue="species")
plt.show()
```



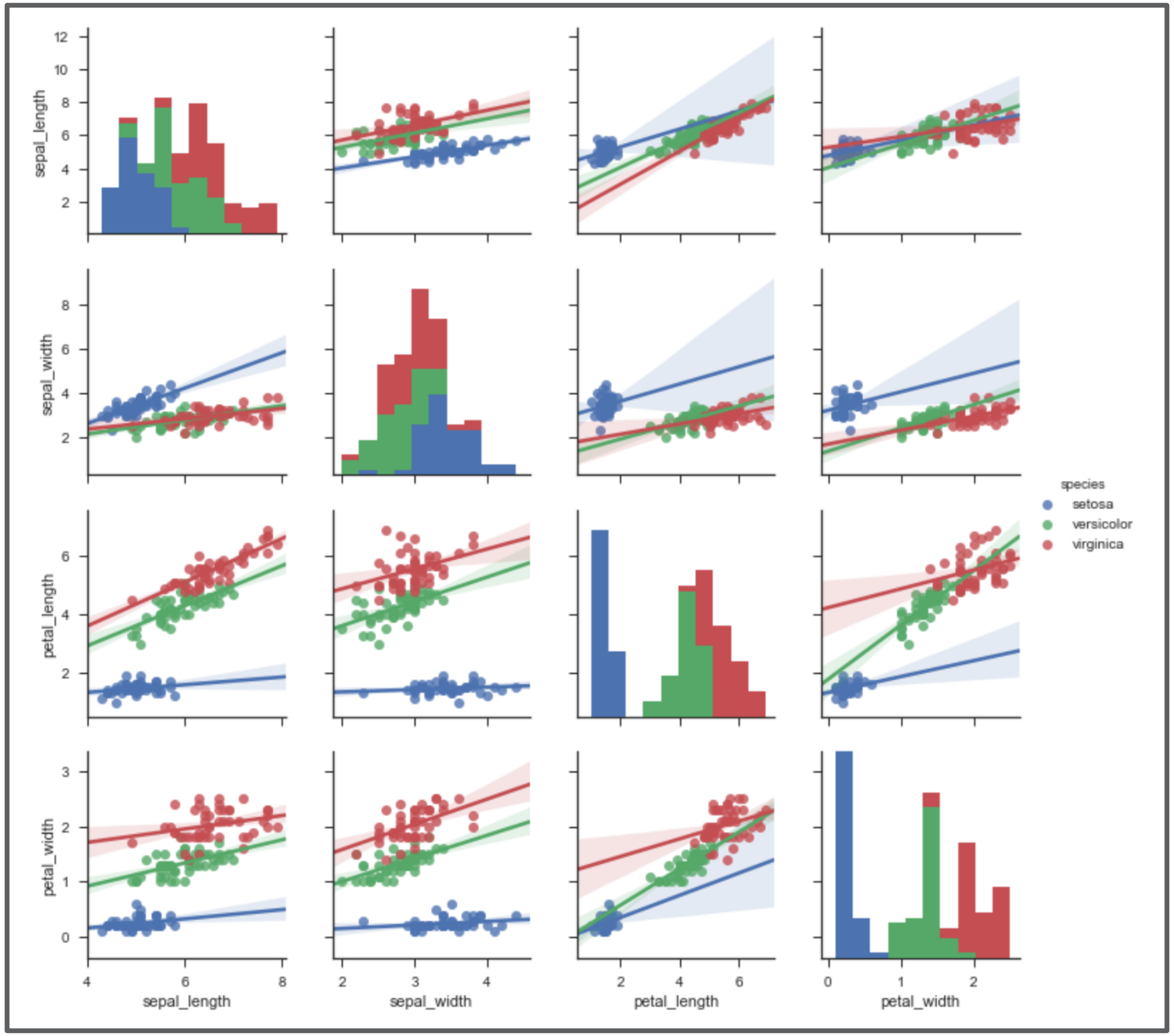
# Visualizing Relationships: Continuous & Categorical Variables



# Visualizing Relationships: Continuous & Categorical Variables

```python
import seaborn as sns
sns.set(style="ticks", color_codes=True)
iris = sns.load_dataset("iris")

sns.pairplot(iris,
             diag_kind='hist',
             hue="species",
             kind="reg")
plt.show()
```



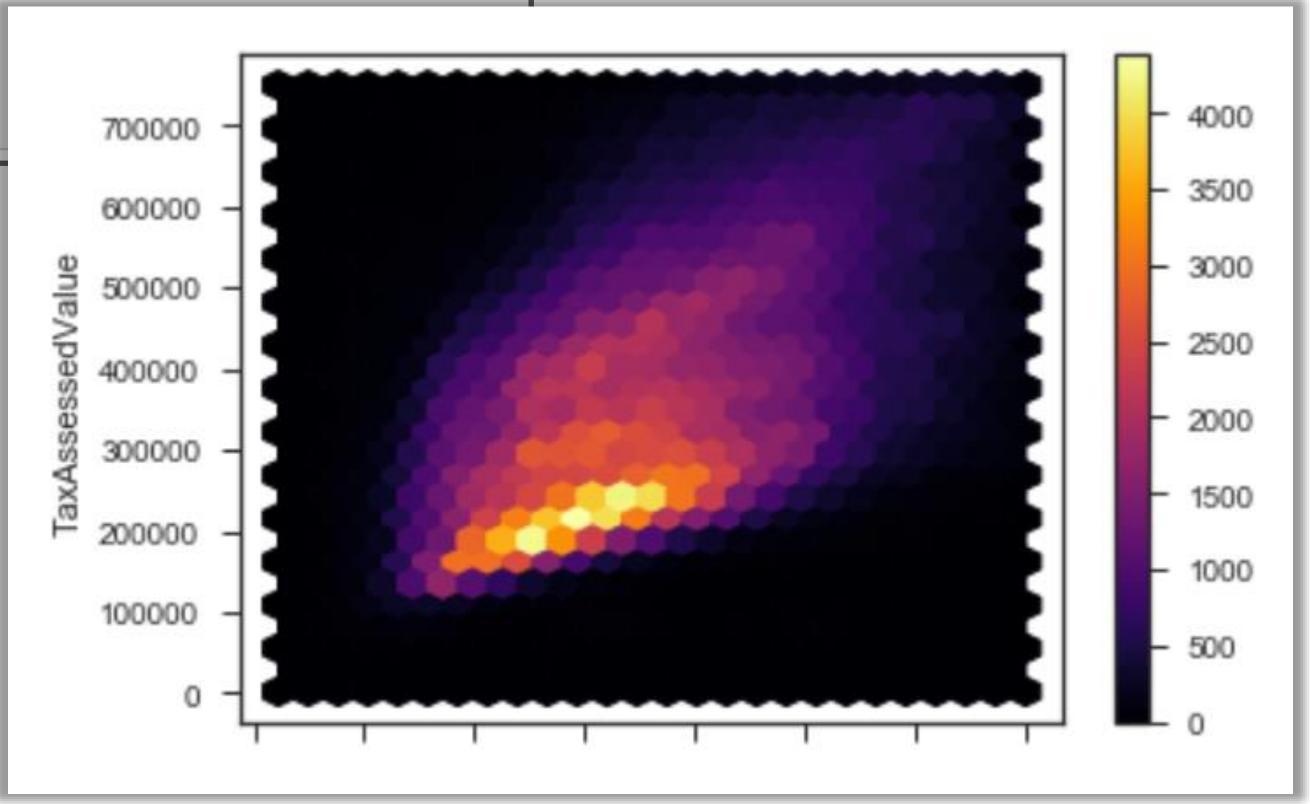
# Hexagonal Binning

## Scatterplots for large-size data

```
kc_tax=pd.read_csv("../data_raw/eda_kc_tax.csv")
kc_tax=kc_tax.query('TaxAssessedValue<750000 & SqFtTotLiving > 100 & SqFtTotLiving < 3500')
kc_tax.shape
```

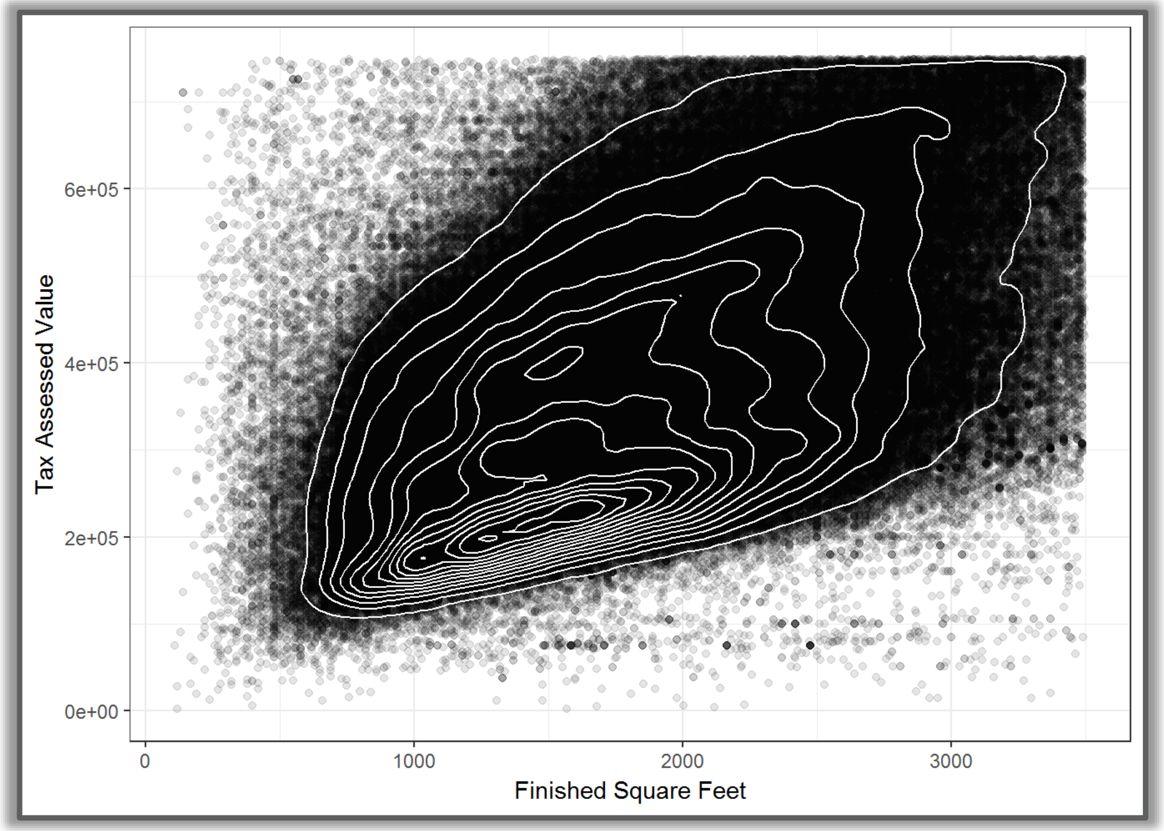
```
kc_tax.plot.hexbin(x='SqFtTotLiving',
                   y='TaxAssessedValue',
                   gridsize = 25,
                   cmap='inferno')
plt.xlabel('Finished Square Feet')
plt.show()
```

Hexagonal Binning



# Contour Plots

## Scatterplots for large-size data



# Quantitative Variable Transformation

## CENTERING AND STANDARDIZING

# Mean

**Let $x = (x_1, x_2, \dots, x_n)$ be the quantitative variable over $n$ observations**

**The mean of the variable:**

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

```
import numpy as np
```

```
x = (2.1, 2.5, 4.0, 3.6)
x_bar = np.mean(x)
x_bar
```

```
3.05
```

$$\bar{x} = \frac{2.1 + 2.5 + 4.0 + 3.6}{4} = 3.05$$

| Economic Growth % ($x_i$) | S & P 500 Returns % ($y_i$) |
|---|---|
| 2.1 | 8 |
| 2.5 | 12 |
| 4.0 | 14 |
| 3.6 | 10 |

# Centering

**Let $x = (x_1, x_2, \dots, x_n)$ be the column: quantitative variable over $n$ observations**

**The mean of the vector:** $$\bar{\boldsymbol{x}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{1}\sum_{i=1}^{n} \boldsymbol{x}_i$$ ← **scalar, number**

**Centering the variable: center $x$ at its mean**

$$\boxed{\boldsymbol{x}_c = \boldsymbol{x} - \bar{\boldsymbol{x}} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})}$$ ← **centered variable**

```python
import numpy as np
```

```python
x = (2.1, 2.5, 4.0, 3.6)
x_bar = np.mean(x)
x_bar
```

```
3.05
```

```python
x_c = x - x_bar
x_c
```

```
array([-0.95, -0.55,  0.95,  0.55])
```

```python
print("{:.2f} : mean of x_c".format(np.mean(x_c)))
```

```
0.00 : mean of x_c
```

**Note: The mean of the centered vector is zero:** $\overline{\boldsymbol{x}_c} = 0$

$$\bar{\boldsymbol{x}} = \frac{\mathbf{2.1 + 2.5 + 4.0 + 3.6}}{\mathbf{4}} = \mathbf{3.05}$$

$$\boldsymbol{x}_c = (\mathbf{2.1 - 3.05, 2.5 - 3.05, 4.0 - 3.05, 3.6 - 3.05})$$
$$= (\mathbf{-.95, -0.55, 0.95, 0.55})$$

| Economic Growth % ($x_i$) | S & P 500 Returns % ($y_i$) |
|---|---|
| 2.1 | 8 |
| 2.5 | 12 |
| 4.0 | 14 |
| 3.6 | 10 |

# Standardizing

| Economic Growth % ($x_i$) | S & P 500 Returns % ($y_i$) |
|---|---|
| 2.1 | 8 |
| 2.5 | 12 |
| 4.0 | 14 |
| 3.6 | 10 |

**Let $x = (x_1, x_2, \ldots, x_n)$ be the column: variable over $n$ observations**

**Centered variable:** $\boxed{\boldsymbol{x_c = x - \bar{x}} = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})}$

**Variance:** $\boldsymbol{var\,(x)} = \dfrac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{\boldsymbol{n-1}}$

**Standard Deviation:** $\boldsymbol{sd(x) = \sqrt{var(x)}}$

**Standardizing using standard deviation:** $\boxed{\boldsymbol{x_s = \dfrac{x}{sd(x)}}}$

**Standardizing using mean & standard deviation ($Z$-score):**

$$\boxed{Z\text{-score} = \boldsymbol{\frac{x-\bar{x}}{sd(x)} = \frac{x_c}{sd(x)}}}$$

# Pair-wise Association of Quantitative Variables

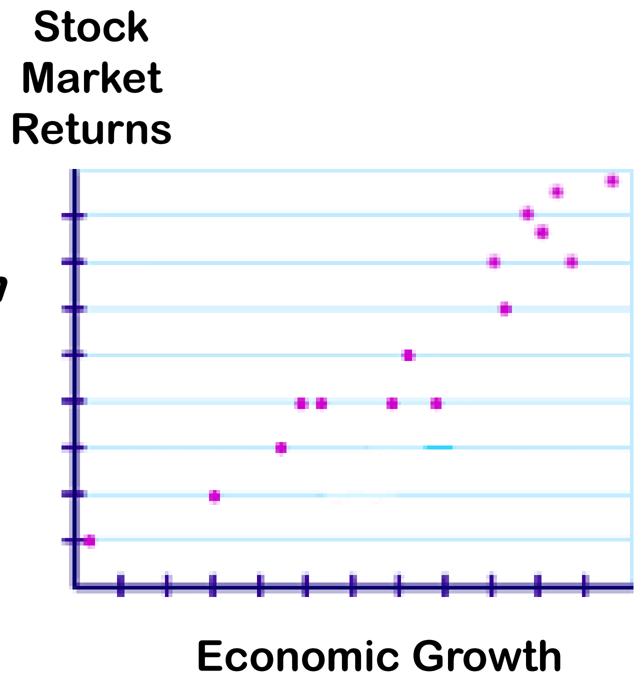
## CORRELATION AND COVARIANCE

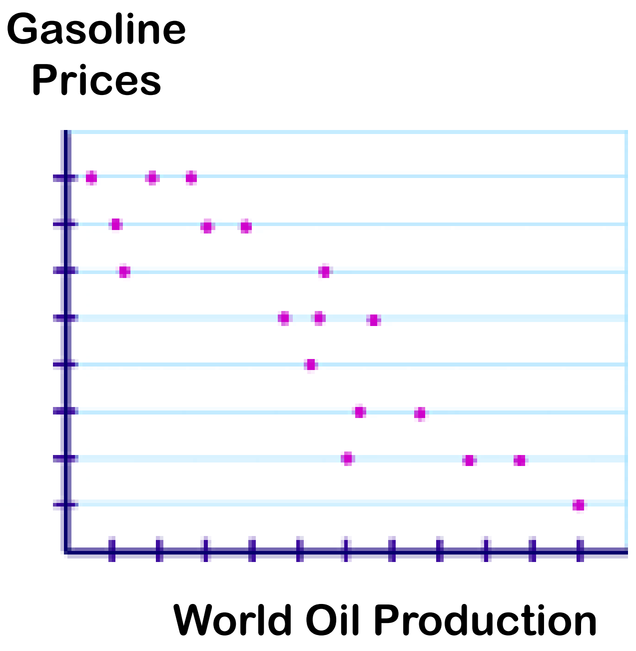
# Covariance and Correlation

Covariance and correlation describe how two quantitative variables are related.

- Variables are positively related if they move in the same direction.
- Variables are inversely related if they move in opposite directions.

*as economic growth increases, stock market returns also increase*




*as oil production increases, gas prices fall*

# Covariance: Formula

$x = (x_1, x_2, \dots, x_n)$ : **the independent variable**

$y = (y_1, y_2, \dots, y_n)$ : **the dependent variable**

$\bar{x}$: **the mean of** $x$

$\bar{y}$: **the mean of** $y$

$n$: **the number of points in the sample**

$$\boxed{cov\,(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}$$

$$\boxed{cov(x, y) = \frac{1}{n-1}\, x_c^T y_c}$$

**Cross-product (inner product) of centered variables normalized by the sample size minus 1 ($n-1$).**

# Covariance: Example

$$cov(x,y) = \frac{1}{n-1}\, x_c^T y_c$$

Cross-product (inner product) of centered variables normalized by the sample size minus 1 ($n-1$).

| Economic Growth % ($x_i$) | S & P 500 Returns % ($y_i$) |
|---|---|
| 2.1 | 8 |
| 2.5 | 12 |
| 4.0 | 14 |
| 3.6 | 10 |

$$cov\,(x,y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}$$

centering

normalized cross-product

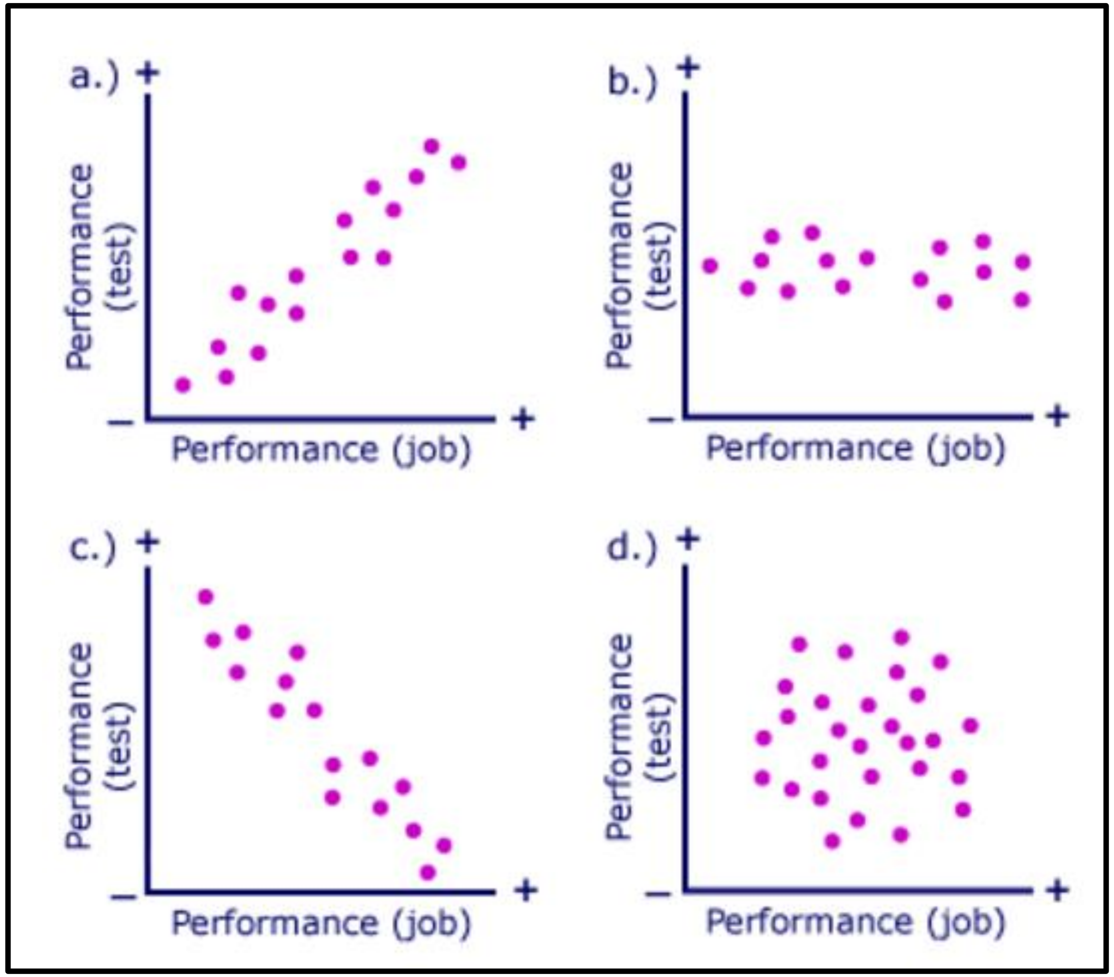
covariance

# Correlation

$$cor(x,y) = \frac{cov(x,y)}{sd(x) \times sd(y)}$$

- **Correlation is a scaled version of covariance (i.e., scaled by the standard deviations)**
- **Correlation and covariance always have the same sign (positive, negative, or 0)**
  - when the sign is positive, the variables are said to be positively correlated
  - when the sign is negative, the variables are said to be negatively correlated
  - and when the sign is 0, the variables are said to be uncorrelated
- **Correlation is dimensionless, since the numerator and denominator have the same physical units.**
- **Correlation will always take on a value between 1 and – 1:**
  - If the correlation coefficient is +1, the variables have a perfect positive correlation. This means that if one variable moves a given amount, the second moves proportionally in the same direction.
  - If correlation coefficient is –1, the variables are perfectly negatively correlated (or inversely correlated) and move in opposition to each other. If one variable increases, the other variable decreases proportionally.
  - If correlation coefficient is zero, no relationship exists between the variables. If one variable moves, no predictions about the movement of the other variable can be made.

# Correlation: Examples




In each of the graphs, are job performance and test performance shown to be positively related, inversely related, or unrelated?

**Answers:**

a) **positively related**
b) **unrelated**
c) **inversely related**
d) **unrelated**

# Exercise: Compute Covariance & Correlation

| Month | Return of Stock A | Return of Market Index |
|---|---|---|
| 1 | 2.3 | 1.3 |
| 2 | 2.5 | 5.0 |
| 3 | 1.9 | 0.8 |
| 4 | 2.4 | 1.9 |
| 5 | 2.1 | 1.1 |

**1. Using the table, show your calculations and Python codes for computing the correlation of Stock A's returns and the return of the market index.**

**2. Do the same for the covariance.**

# Exercise: Solution

| Month | Return of Stock A | Return of Market Index |
|---|---|---|
| 1 | 2.3 | 1.3 |
| 2 | 2.5 | 5.0 |
| 3 | 1.9 | 0.8 |
| 4 | 2.4 | 1.9 |
| 5 | 2.1 | 1.1 |

| | | | step 1 | step 2 |
|---|---|---|---|---|
| | Stock A | Market Return | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
| | 2.30 | 1.30 | 0.0036 | 0.5184 |
| | 2.50 | 5.00 | 0.0676 | 8.8804 |
| | 1.90 | 0.80 | 0.1156 | 1.4884 |
| | 2.40 | 1.90 | 0.0256 | 0.0144 |
| | 2.10 | 1.10 | 0.0196 | 0.8464 |
| Sum | | | 0.2320 | 11.7480 |
| Average | 2.24 | 2.02 | | |
| Sum ÷ 4 | | | 0.0580 | 2.9370 |
| Standard deviation | | | 0.2408 | 1.7138 |

$$cor(x,y) = \frac{cov(x,y)}{sd(x)sd(y)} =$$

$$= \frac{0.31}{(0.24)(1.71)} = \frac{0.31}{0.41} = 0.76$$

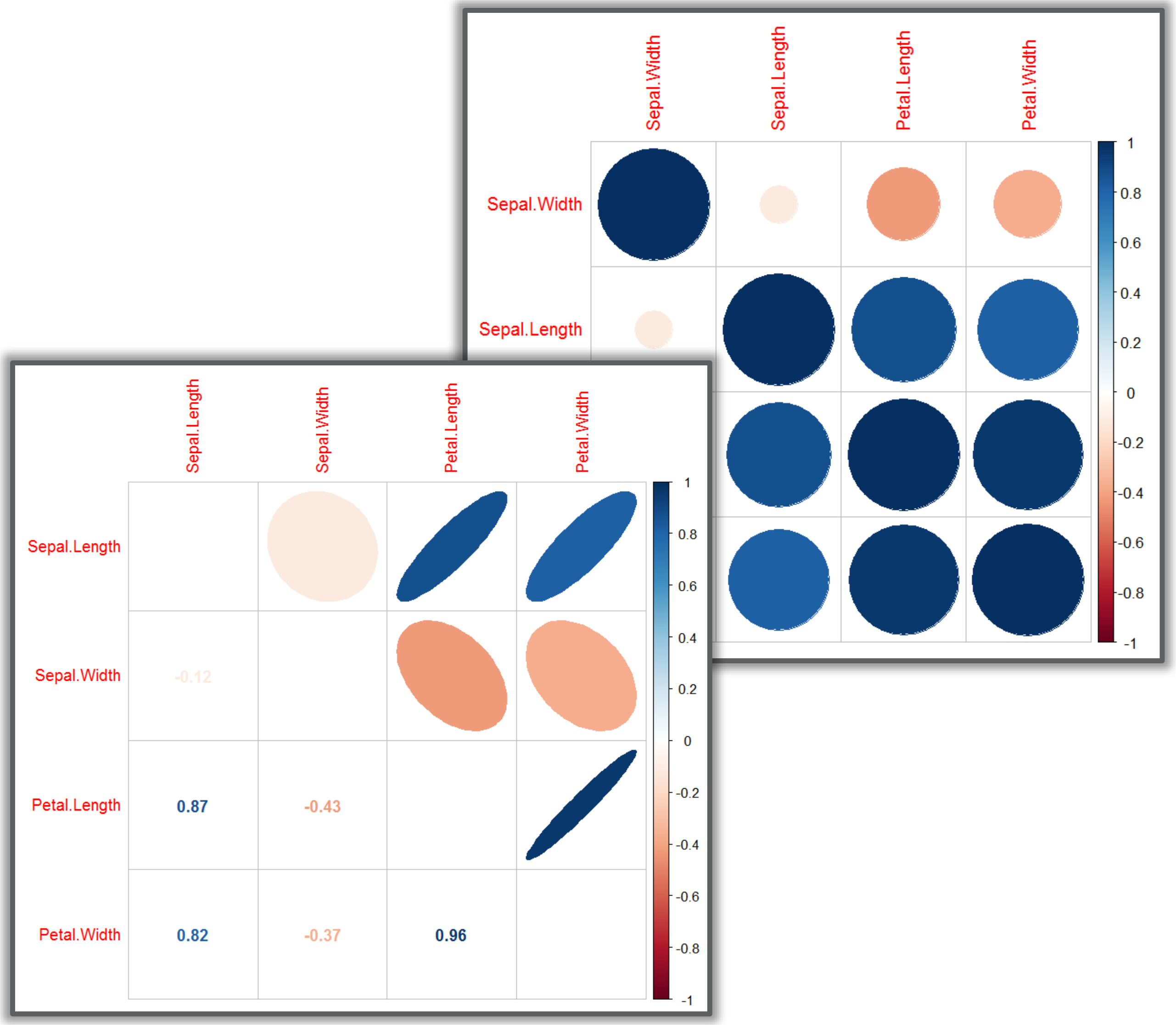
# Correlation Matrix: Pairwise Correlations: Visualization

```python
import seaborn as sns
df = sns.load_dataset("iris")
df.head()
```

```python
df_corr_matrix = df.corr()
df_corr_matrix.head()
```

| | sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|---|
| **sepal_length** | 1.000000 | -0.999205 | 0.999661 | 0.999485 |
| **sepal_width** | -0.999205 | 1.000000 | -0.999904 | -0.999969 |
| **petal_length** | 0.999661 | -0.999904 | 1.000000 | 0.999982 |
| **petal_width** | 0.999485 | -0.999969 | 0.999982 | 1.000000 |

```python
from biokit.viz import corrplot
c = corrplot.Corrplot(df_corr_matrix)
c.plot()
```



# Correlation: Statistical Significance

```
import seaborn as sns
df = sns.load_dataset("iris")

from scipy import stats

pearson_coef, p_value = stats.pearsonr(df["petal_length"],
                                      df["petal_width"])
print("Pearson Correlation Coefficient: ", pearson_coef,
      "\nand a P-value of:", p_value)
```

```
Pearson Correlation Coefficient:  0.962865431403
and a P-value of: 4.67500390733e-86
```

**$p$-value < 0.05: statistically significant**

# Other Resources

- https://seaborn.pydata.org/generated/seaborn.pairplot.html
- https://machinelearningmastery.com/visualize-machine-learning-data-python-pandas/
- https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166
- http://thomas-cokelaer.info/blog/2014/10/corrplot-function-in-python/