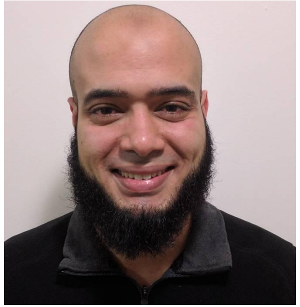
Machine Learning

Linear Regression Assumptions

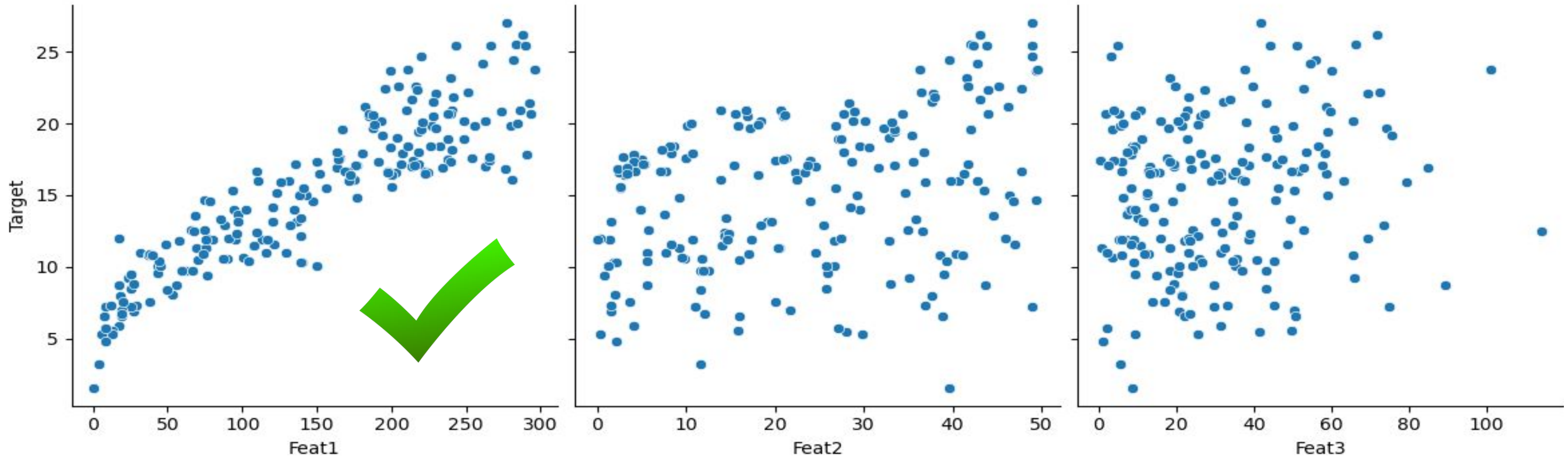
Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



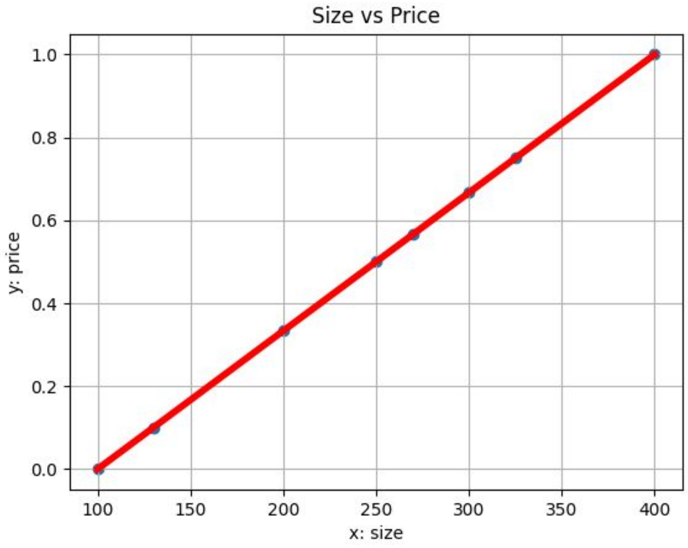
Recall the homework

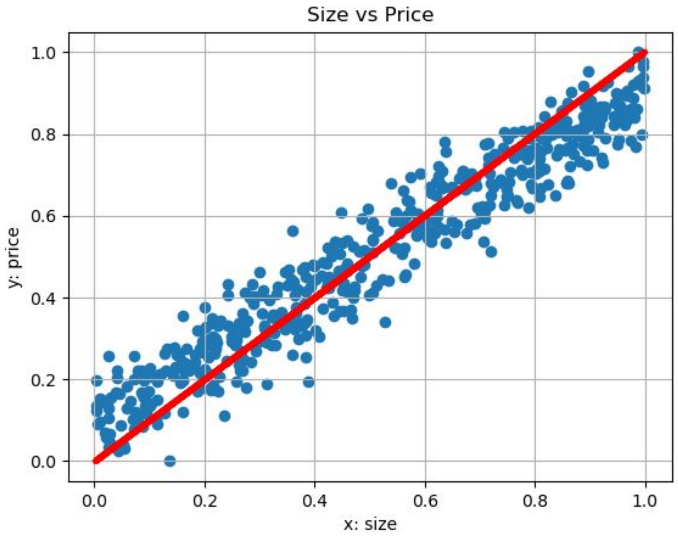
- When we visualized the 3 features separately, we discovered that only one of them that seems to generate a line, **while the other 2 features can't be modeled using a line!**



Recall these 2 plots

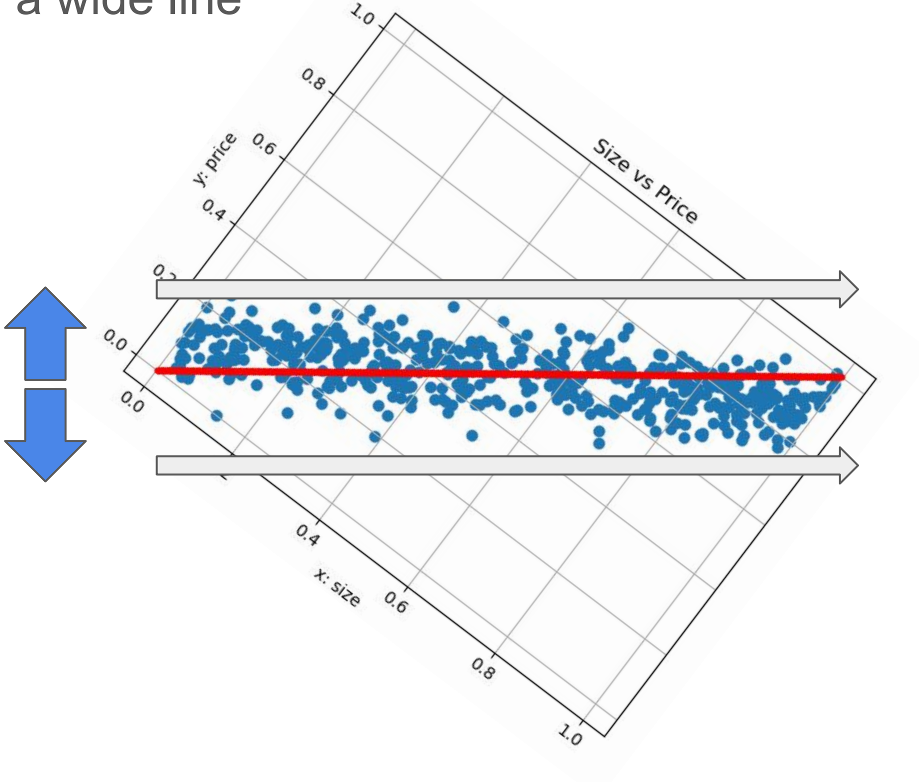
- Left: Points generated on the line: y = 3*x - 50
- Right: Points generated on the line: y = 3*x - 50 + **noise**
  - Noise is sampled from **normal** distribution (mu=0, sigma=0.3)





Observe

- Visually, the data seems to be creating a wide line
    - We call that **Linearity**
- With increasing x, the points have similar variance around the line
    - We call that **Homoscedasticity**
        - Aka Constant Variance
    - See the 2 long left-to-right arrows
- Around each y, the points are above and below each y in a symmetric way
    - We call that **Normality**
    - See the blue up/down arrows

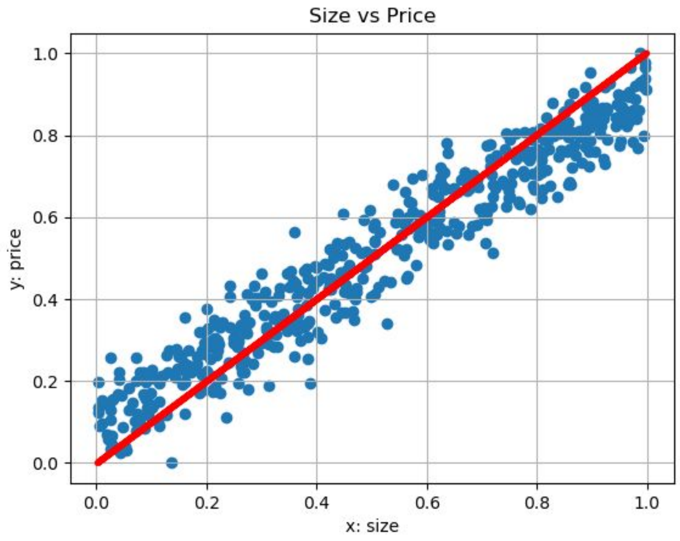


Linear Regression **Assumptions**

- How can we **trust** the linear regression model's predictions?
- Statisticians identified several **factors** such that the **data can be modeled** using linear regression
  - Their violations might have consequences
- Different procedures were identified to verify the assumptions
  - There are libraries for that. Beyond the scope of this course
- *Interviewers sometimes ask you to state/explain the assumptions*

Linearity **Assumption**

- **Linearity** means that there must be a **linear relationship** between the **independent** (feature) variable and the **dependent** (target) variable
- What if the assumption is NOT satisfied?
  - The model won't work!
  - There can be some features that are not linear but at least enough features are required for some good performance



Question!

- Assume we have 2 parameters to learn (m and c)
- Input is x and output is y
- Which of the following is a linear equation, relative to the parameters?
  - $y = mx + c$
  - $y = mx^2 + c$
  - $y = mx + x^3 + c$

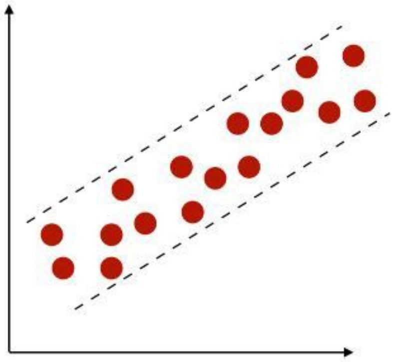
- All of them are linear in **m and c**.
- x and x^3 are just coefficients!

Homoscedasticity Assumption

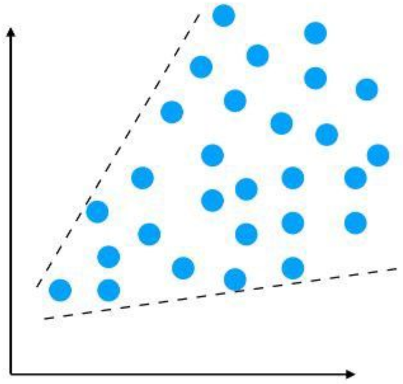
- The **residual** errors should have constant **variance**
- What if the assumption is NOT satisfied?
  - It's unlikely to be a problem for **predictions** on the line
  - Some statistical tests/values will be wrong (f-test, t-value, confidence intervals, etc)
  - So we have 2 concerns:

    1) Prediction

    2) Statistical tests



Homoscedasticity



Heteroscedasticity

Img src

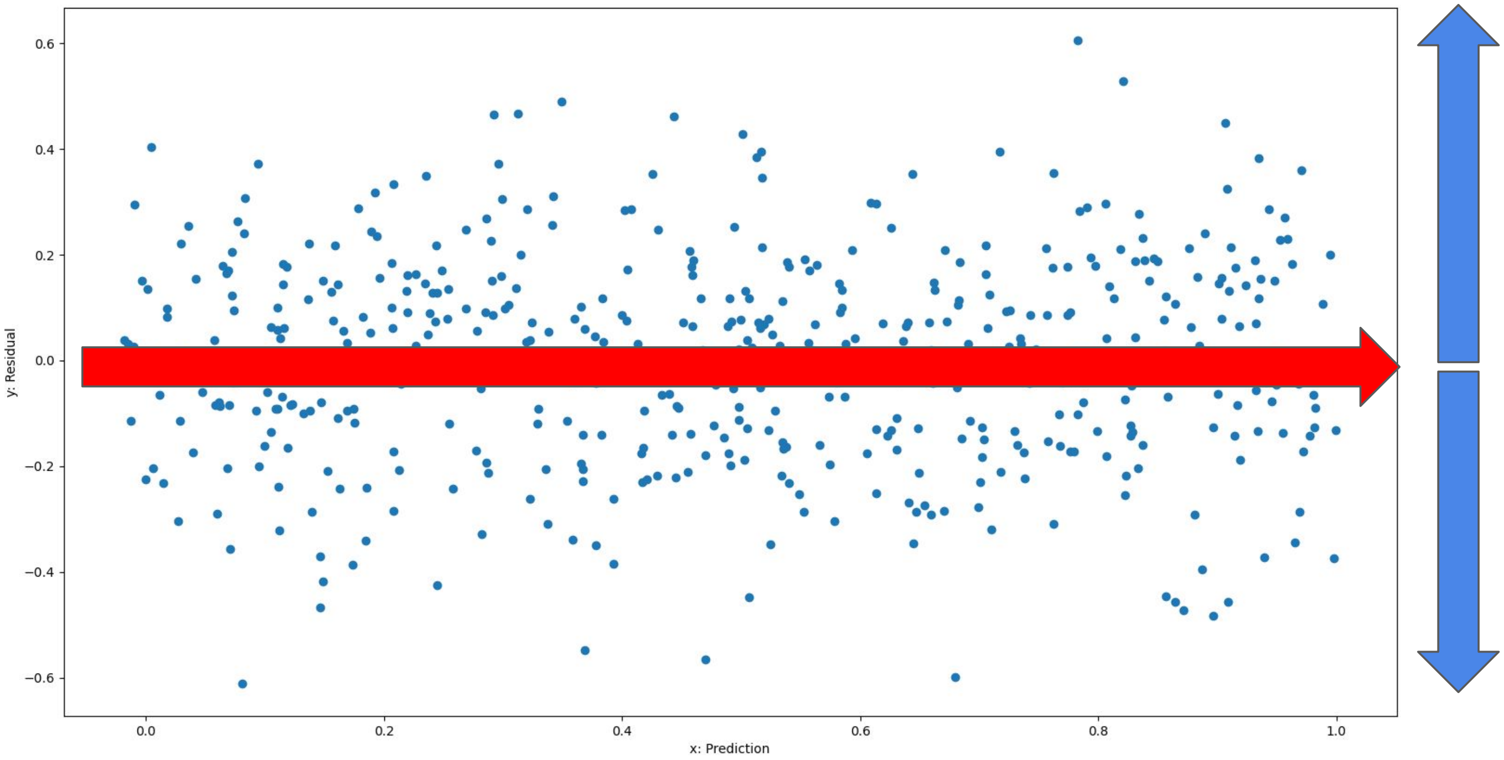
Residual Plot for Verifications

- Plotting residuals may help us **verify** the previous 2 assumptions
- We plot the residual values on the y-axis vs **predicted**-value on the x-axis
  - Why not vs the x itself?
    - Real-life problems are not based on a single variable (easy 2D plot)
  - Both residual and predicted values are just **scalars,** even for multivariate input
  - How to get residuals?
    - 1) Compute the model
    - 2) Do predictions
    - 3) Compute error (residuals)

Residual Plot: Code

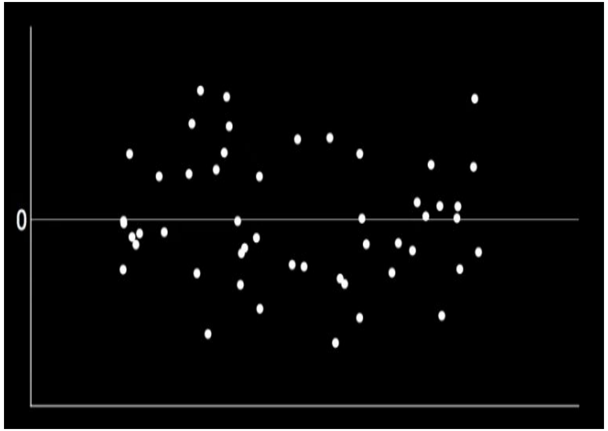
```python
import matplotlib.pyplot as plt
import numpy as np
from sklearn import linear_model

x = np.random.rand(500)
noise = np.random.normal(0, 0.2 , 500)
t = x + noise   # diagonal line
x, t = x.reshape(-1, 1), t.reshape(-1, 1)

pred_t = linear_model.LinearRegression().fit(x, t).predict(x)
residuals = t - pred_t
plt.scatter(pred_t, residuals)
plt.xlabel('x: Prediction')
plt.ylabel('y: Residual')
plt.show()
```

Residual Plot: Visualization

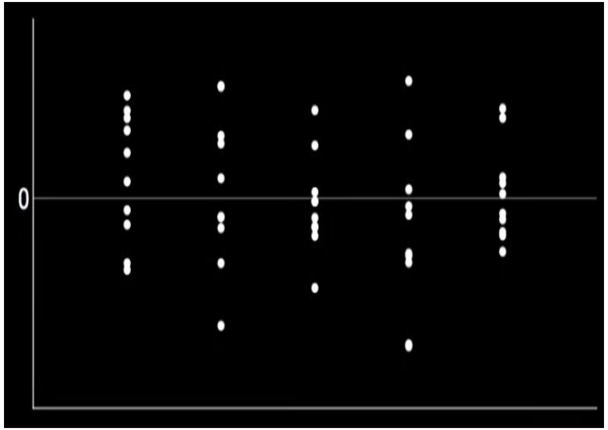


Question!

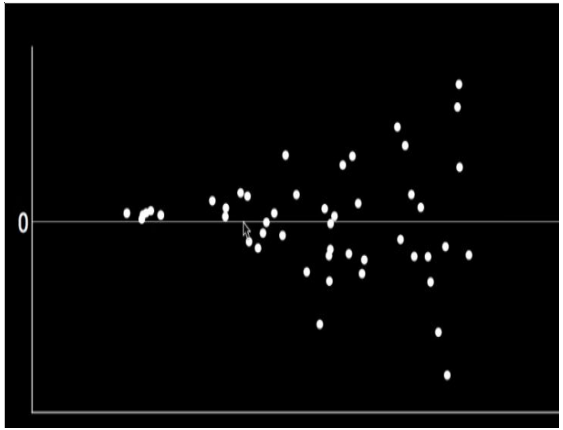
- Which of the following residual plots satisfy the 2 conditions?



A



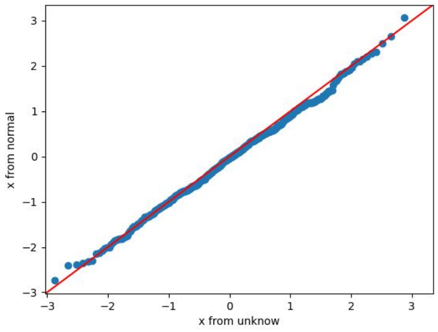
B

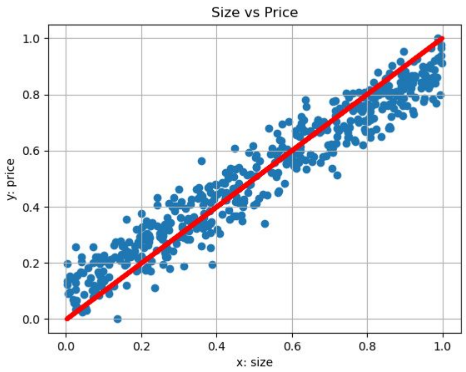


C

Normality **Assumption**

- Normality means that the residuals (target - prediction) that result from the linear regression **model** should be **normally** distributed
- What if the assumption is NOT satisfied?
  - Sometimes **outliers** are the reason. Remove them first
    - An outlier is an **unusual** data point that differs **significantly** from other data points
  - Again, some statistical tests will not be reliable
- As we learned, we can use a QQ plot to verify the residuals from the normal distribution





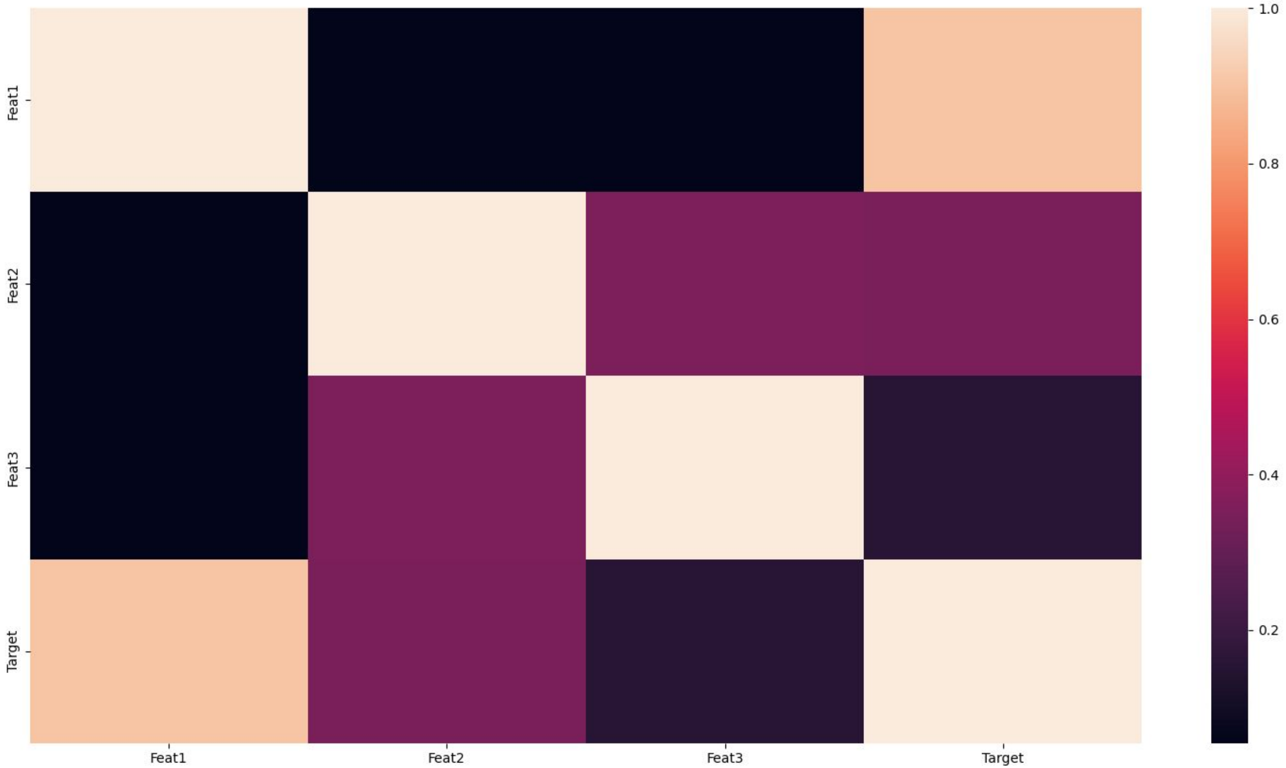
No Multicollinearity **Assumption**

- Multicollinearity happens when independent variables in the regression model are highly correlated to each other
    - Assume we have 4 features: birth_day, birth_month, birth_year, total_days
    - total_days can be **derived** from the first 3 features
    - But the regression model is trying to find an **independent** coefficient for it!
- How to handle?
    - Check **correlation** between variables
    - Build a regression **model** to regress one feature from the remaining
    - Find and remove problematic features

Correlation

- **Correlation** is a measure that expresses the extent to which two variables are linearly related (changing **together** at a constant rate)
  - **Positive** Correlation: increase together and decrease together
  - **Negative** Correlation: When one increases and the other decreases
  - **Zero**/No Correlation: seemingly unrelated variables
  - A correlation is usually tested for two variables at a time, but you can test correlations between *three or more variables*.
- Correlation **matrix**: calculates the **pairwise correlation** coefficients between two or more variables
  - Algorithms for the computation: **Pearson** (most common), Kendall, Spearman
  - We can also convert the matrix to a **heatmap**
  - A **heat map** is a 2D representation of data in which values are represented by **colors**

	Feat1	Feat2	Feat3	Target
Feat1	1.000000	0.054809	0.056648	0.901208
Feat2	0.054809	1.000000	0.354104	0.349631
Feat3	0.056648	0.354104	1.000000	0.157960
Target	0.901208	0.349631	0.157960	1.000000



The last column is vital for showing the linearity of features with the target

Ignore the diagonal

```python
df, _, x, t, _ = load_data('data/dataset_200x4_regression.csv')

# You can use pandas to get the correlation matrix
correlation_matrix = df.corr()
round(correlation_matrix, 2)
print(correlation_matrix)

# plot the matrix heatmap
sns.heatmap(correlation_matrix)
plot.show()

# we can also use compute correlation of 2 columns of data
# using pandas
ans = df['Feat1'].corr(df['Target'])    # panda series
print(ans)
# using stats
ans = stats.pearsonr(df['Feat1'], df['Target'])[0]
print(ans)

ans = stats.pearsonr(x[:, 0], t)[0]
print(ans)
```

Correlation does not imply causality

- Important concept in statistics and real life!
- The data suggests that **Ice Cream Sales and Shark Attacks** have positive correlation
    - So, people eating more ice cream causes more shark attacks? Absolutely not!
    - Ice cream is usually sold during the summer, and it is during the summer that people are more likely to go swimming.
    - The increased shark attacks are simply caused by **more time in the water**, not ice cream
- More Examples
    - Master's Degrees vs. Box Office Revenue
    - Pool Drownings vs. Nuclear Energy Production.
    - Measles Cases vs. Marriage Rate
    - High School Graduates vs. Pizza Consumption

Question!

- Assume that $y = 2x_1 + 5x_2 + 4x_3$
- Are x_1 and x_3 are correlated if the following is true:

  - 1) $x_3 = 5x_1 + 7$

  - 2) $x_3 = 5x_1^2$

- 1) Yes, clear linear relationship
- 2) No, the relation in our definition is linear. This is quadratic

Relevant Resources

- Simple Linear Regression: Checking Assumptions with Residual Plots
- Assumptions of Linear Regression - DATAtab **channel**
- Assumptions of Linear Regression - Article
- Assumptions of Linear Regression - Article
- Regression assumptions explained - zedstatistics **channel**
- Multicollinearity in Regression
- What Happens When You Break the Assumptions of Linear Regression?
- A Critical and Often Misunderstood Fact About Linear Regression
- Why normality of residuals barely important for estimating the regression line?

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."