Why Pre-Process Data?

- Typical **problems** with the data include: missing values; inconsistent data; outliers; skewed or bi-modal distributions.
- Some times the data is simply not in the **appropriate format** for the type of analysis to be done.
- Some times the data needed for the analysis needs to be gathered from multiple data sources and joined
- Or the data needs to be **scaled**, **centered** or **normalized**
- Or it needs to be aggregated, summarized, etc.
- Especially when data is not at the **"unit of analysis"** level
- For example, you may have data on prices and features for individual houses, but you need to analyze housing by counties in the US – i.e., the unit of analysis is the **county**, not the individual homes → the data needs to be **aggregated**

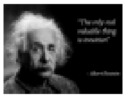
TRANSFORMATION 1: Categorical data to dummy variable predictors

KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

The Dummy Variable Trap

- This is a well-known problem when you convert a categorical variable into various **“mutually exclusive”** dummy variables.
- For example, if you have a categorical variable called “LocationType” and it has one of **three possible values** (Urban, Suburban and Rural) we can create 3 dummy variables called Urban, Suburban and Rural, respectively.
- If LocType = “Urban” → **Urban = 1**; 0 otherwise
  If LocType = “Suburban” → **Suburban = 1**; 0 otherwise
  If LocType = “Rural” → **Rural = 1**; 0 otherwise
- However, these three dummy variables are **mutually exclusive**, so if Urban = Suburban = 0, then Rural must be 1.
- That is, the value in any of these variables is fully dependent on the other 2
- Including all 3 variables in a regression model will not only violate the **assumption of independence**, but will also create **infinite multicollinearity** and **infinite standard errors**

TRANSFORMATION 2: Polynomials

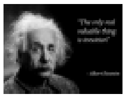


KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

Polynomials: Intuition

- Polynomial transformations are very useful when the relationship between the X's and Y are suspected to be **non-linear**
- We cover **non-linear models** in depth later on, so we will only discuss this briefly here
- Generally, a quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ is the preferred polynomial if the data is **curvilinear** or **with 1 peak/valley**.
- A **cubic** model is preferred with **2 peaks/valleys**, etc.
- The more **wavy** the relationship the higher the polynomial
- The problem is that **high polynomials** are **difficult to interpret** and they tend to **"over-identify"** the model and do not generally perform well with new data, especially at **both ends** of the curve
- **Spline** and **piecewise** models (covered later in the class) generally perform better than high polynomials.
- Quadratic and cubic transformations are the **most popular** polynomials.

TRANSFORMATION 3: Box-Cox



KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

Box-Cox: Intuition

- Box-Cox is a **family** of transformations for the response variable y^*
- You get a different transformation y^* for every value of **λ** in
  - For $\lambda = 0 \rightarrow y^* = Log(y)$
  - For $\lambda \neq 0 \rightarrow y^* = \frac{y^{\lambda} - 1}{\lambda}$
- Box-Cox transformations are useful when one is having **difficulties** obtaining a transformed variable with a **normal distribution**.
- The idea is to **systematically** calculate y^* for **λ = 0, 1, 2, etc.** and select the transformation that yields the most normally distributed y^*
- Box-Cox transformations are a useful statistical **feature engineering** technique, but the transformed variables are difficult to interpret. However, they may prove to be very useful when **predictive accuracy** is a more important goal than interpretation

TRANSFORMATION 4: Log Models

KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

Log Models: Interpreting Effects

A model $y = \beta_0 + \beta_1 x + \varepsilon$ has 4 possible linear-log models
(more than one x variable can be logged):

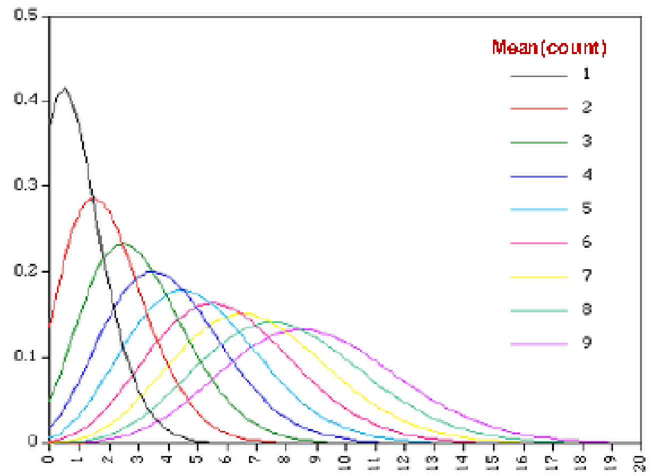
Dependent Variable	Independent Variable	
	x	Log(x)
y	**Linear Model** $y = \beta_0 + \beta_1 x + \varepsilon$ x ↑ 1 unit → y ↑ β_1 units	**Linear-Log Model** $y = \beta_0 + \beta_1 Log(x) + \varepsilon$ 1% (1/100) ↑ in x → y ↑ β_1/100 units
Log(y)	**Log-Linear Model** $Log(y) = \beta_0 + \beta_1 X + \varepsilon$ x ↑ 1 unit → y ↑ 100*β_1% (i.e., β_1 fraction)	**Log-Log (Elasticity) Model** $Log(y) = \beta_0 + \beta_1 \mathrm{Log}(x) + \varepsilon$ 1% ↑ x → y ↑ β_1%

TRANSFORMATION 5: Count data

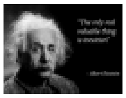
KOGOD SCHOOL of BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

The Poisson Distribution

- The **Normal** distribution: is **symmetrical** around the mean; it's **tails** extend **indefinitely** at both ends; and it is **continuous**.
- The **Poisson** distribution: is bounded at **0**; is **asymmetrical**; **varies** in **shape** as the mean increases (it approaches a normal distribution when counts have very wide ranges)
- As it turns out, **count data** follow a **Poisson** distribution
- Notice in the diagram how the **shape** of the distribution curve **changes** with the **count mean** and how it becomes **more normal** with **large** means.



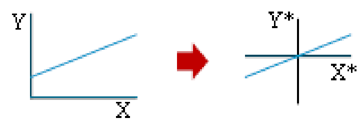
TRANSFORMATION 6: Centering data



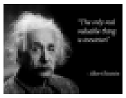
KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

Centering: Intuition

- There are times when centering one or more variables around their means is useful.
- For example, maybe the **intercept is meaningless** because **x cannot be 0** (e.g., predicting cholesterol based on weight → nobody has 0 weight). Centering the variable shows the effect of x when x is at its mean value (not zero).
- Also, when computing **interaction** variables of **2 continuous** variables x_1 and x_2 , the resulting interaction term $x_1 * x_2$ is **problematic** for a number of reasons:
  - **Scale invariance** – changing the scale of x_1 or x_2 (e.g., from feet to meters) will change the main effect size
  - The product of $x_1 * x_2$ may generate severe **multicollinearity**
- **Centering** x_1, x_2 and y with respect to their means helps → $x_1^* = x_1 - \bar{x}_1$; $x_2^* = x_2 - \bar{x}_2$
- This is **equivalent** to **shifting** the Y and X **axes** to the Y and X means



TRANSFORMATION 7: Standardizing data



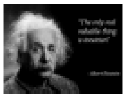
KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

Standardization: Intuition

- It is sometimes useful to **divide** a **centered** variable by the variable's **standard deviation**
- This produces a transformed variable with $\bar{x} = 0$ and $\sigma = 1$, often called a **"standard score"** or **"z-score"**: $y^* = \frac{y-y}{\sigma_y}$ $x^* = \frac{x-x}{\sigma_x}$
- This is very useful when you want to **compare dissimilar scales** (e.g., is weight larger than height?) or when the **effect size** of an unstandardized variable has **no meaning**.
- For example, in **survey** studies, we often see rating questions (e.g., rate your satisfaction from 1 to 7). So, what is the meaning of the effect from increasing the response by 1 scale point (e.g., 4 to 5)? It has no meaning.
- Standardizing x and y in $y^* = \beta_0 + \beta_1 x^* + \varepsilon$ the **interpretation** is:

  x ↑ 1 standard deviation → y ↑ β_1 standard deviations
- **Fun fact:** in a simple regression model like the one above, the resulting **standardized coefficient** is identical to the **correlation** between y and x

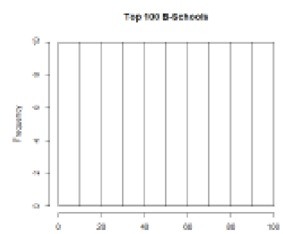
TRANSFORMATION 8: Rank transformations



KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

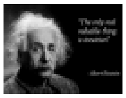
Rank: Intuition

- Sometimes a variable is important but does not have **enough variance** to capture significant effects in regression models
- Sometimes a variable's **distributions** is **uneven** or non-normal, which is particularly problematic for **small samples**
- A **rank transformation** is a popular **"non-parametric"** statistical approach, which solves some of these problems
- This is done by **sorting** the values and assigning **1** for the smallest value, **2** for the next, etc., or vice versa (highest to smallest)
- The intervals between data points is exactly 1 (i.e., the next value after rank 1 is rank 2), rank transformation has the nice property of producing a **"uniform distribution".** For example the ranks of the top 100 B-Schools distribution looks like this



- **Effect interpretation** → the unit increase or decrease is the rank, not the actual value (e.g., increase rank by 1)
- Some times ranks are **re-scaled** to a **0-1** scale

TRANSFORMATION 9: Lagging data



KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

Lagged Models: Intuition

- Lagged models are popular in predictive models with an **ordinal** or **sequence** variable (e.g., **time**, distance)
- We will focus on **"time"** as the ordinal variable, but the principles apply to any other ordinal variable
- **Time-based** models come in different flavors of **two main types**
  - **Time Series** – predicting future values of a variable based on its prior values (e.g., predict tomorrow's weather from today's)
  - **Causal Models** – like time series, but also include other **predictors** and **control** variables
- Causal models are more useful in predictive modeling because one can **control for** various factors that may influence results
- The name "causal" is misleading because statistical **correlation** does not imply **"causality"**.
- **Lagging** some predictors provide **stronger causal models** because we use past values to predict future ones
- Lagged variables often help correct **serial correlation** problems too