# Feature Engineering

## Methods

- Recategorize
- Infer feature from another variable
- Dimension Reduction
    - PCA
    - Cluster Analysis
- Apply Descriptive Statistic
    - Central tendency
    - Variability
    - Extremes
    - Trends

## Common Mistakes and Limitations

- Using the outcome variable
- Not using descriptive statistics properly
- Too much overlap with other predictor variables
- Multicollinearity

**Note-** Cross validation is the best way to determine whether to keep or shelf a featured variable. If the cross validated error improves with the feature added, we should keep it.

# Example-

| Participant | Purchases | Credit Card Zipcode | Age | Married | Phone OS | Preferred Shipping Times | Purchases 1 month ago | Purchases 2 months ago | Purchases 3 months ago |
|---|---|---|---|---|---|---|---|---|---|
| Aaron | 6 | 10013 | 29 | Yes | iPhone | After 5pm | 4 | 2 | 0 |
| Brenda | 5 | 10065 | 30 | No | Pixel | After 5pm | 4 | 0 | 2 |
| Carly | 2 | 10451 | 27 | Yes | Cricket | 2pm | 2 | 3 | 4 |

With feature engineering, we can add on these four additional rows based on the original answers provided

We can create these new variables from the information we have in the original variables

| Participant | Purchases | Credit Card Zipcode | Age | Married | Phone | Preferred Shipping Times | Purchases 1 month ago | Purchases 2 months ago | Purchases 3 months ago | Probable Household Income | Probably Employed | 3 month Trajectory | 3 month Variability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aaron | 6 | 10013 | 29 | Yes | iPhone | After 6pm | 4 | 2 | 0 | $83,580 | Yes | 2 | 2 |
| Brenda | 8 | 10065 | 30 | No | Pixel | After 7pm | 6 | 3 | 0 | $102,883 | Yes | 3 | 2 |
| Carly | 2 | 10451 | 27 | Yes | Cricket | Between 9am and 3pm | 2 | 3 | 4 | $29,043 | No | -1 | 1 |

Median income for the zipcode listed

Based on hours available at home

The average increase per month, and mean absolute deviation

# Feature Engineering Methods

## How to Engineer Features

- **Recategorize:** Convert into new categories based on cutoffs
- **Dimension reduction:** Find the underlying groups and factors within a collection of responses
- **Apply descriptive statistic:** Represent a collect of attributes in terms of a summary statistic
- **Infer feature from other variables**: Use collection of variable values to infer other attribute

# Recategorization

**Recategorization** is when you group certain values into a cohesive category

- Typically accomplished by:
  - setting cutoffs between ranges of values
  - assigning anyone in certain categories to another category
- There are many ways to categorize/group the values
- Promotes finding relationships between predictors and outcomes by explicitly indicating socially meaningful groupings

# Examples of Recategorization

**Recategorize Categorical Variables**

- Monday, Tuesday, Wednesday, … becomes Weekend vs. Weekday
- Heat with Humidity becomes Sweltering

**Recategorize a date:**

- 11:00 3:00 7:00 … becomes morning, afternoon, evening
- Separate time into before work hours and after work hours

**Recategorize Ratio Variables**

- 5-yrs to 12 is child, > 65 is retired

# Inferring Attributes

Similiar to how model based missing data methods can infer an attribute based on the values of the other variables, you can infer people's attributes based on the combination of responses provided.

Works only if you combine the information from multiple variables or from a nominal meta variable (zipcode, license#). Otherwise, we are just duplicating the information from another variable

**Example:**

- We have a person's zipcode, so we can probabilistically infer the person's household income
- We know this person is never active on the site during work hours, so we can probabilistically infer this person is employed
- We know this person is 6'7 and weighs 250 lbs, so we can probabilistically infer that this person is male

# Examples

**Use demographic tendencies**

- People with these last names tend to be Hispanic
- People with these heights and weights tend to be male

**Use behavioral tendencies**

- People with this activity level during the day and this level of education tend to be unemployed
- People who have seen these movies tend to be female
- People with this writing style tend to be educated

**Match an entry with archival records**

- This day was sunny and 90 degrees
- This customer's LinkedIn says their educational background is a Bachelor's degree
- This customer's zipcode has a median income of $50,000

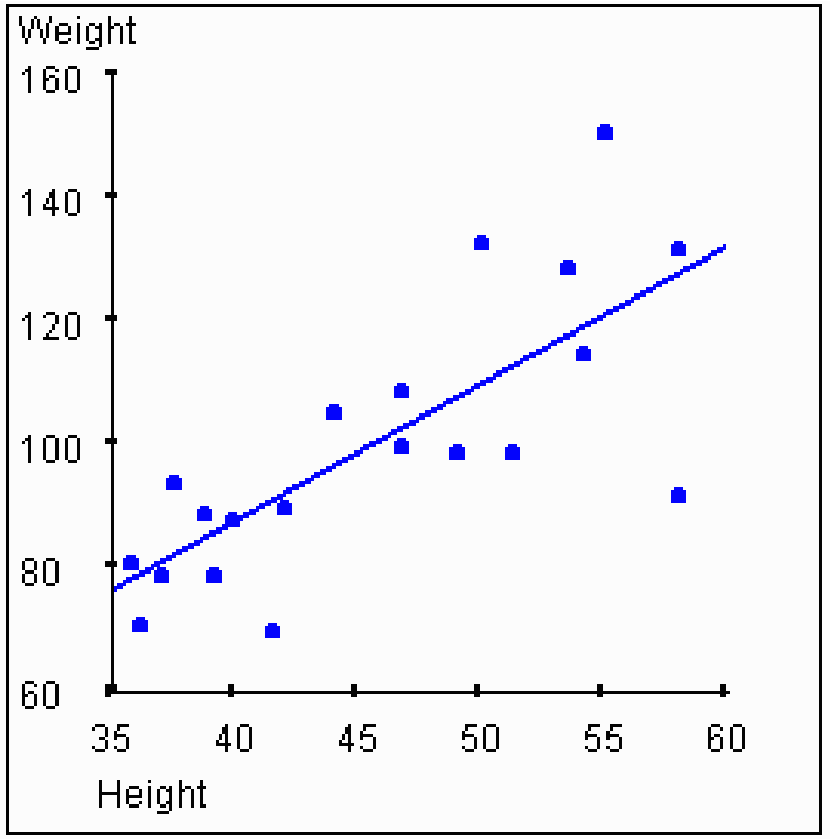
# Common mistakes during FE

- ✓ Using the outcome variable
- ✓ Not using descriptive statistics appropriately
- ✓ Too much overlap with other predictor variables
- ✓ Not enough creativity
- ✓ Spending too little time

# Linear Regression

- If the relationship between two variables fits a straight line, then their values are very predictable

- We predict the outcome of a variable that is linearly related to another variable by using **Linear Regression**




THE OBJECTIVE OF REGRESSION IS TO FIND THE LINE OF BEST FIT

# Regression Algorithm Steps

**Linear Regression test process**

- 1)  Pick a research question that identifies the population, an outcome, and a predictor
- 2) Collect a sample of data from the population
- 3) Check the assumptions
- 4) Run the regression analysis
- 5) Examine the slope and intercept
- 6) Examine the Proportion of Variance Explained in the Outcome ($R^2$)
- 7) Interpret the results
- 8) Follow up the results with a prediction