# **Week 10 missing values (Imputation); Optimization modes**

## Agenda:

Prescriptive analytics

Optimziation models

Advanced data preparation

# 10.1 Missing data

**Assumption:**

1. lot of data

Data point(i) is in millions

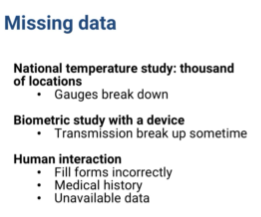
Factors (j) is in 1000s

1. Assume we know all the data

****

**Agenda:** what if some data is missing

### Example:

****

* Wifi could be broken if biometric reading

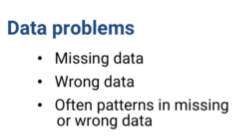
-doctor form filling

-leave telephone number blank

-application filling

-purchase data set can be missing data

### Ways of missing data

****

* if data is so wrong, it is outlier

patterns:

1. some pieces are more likely to be missing than others

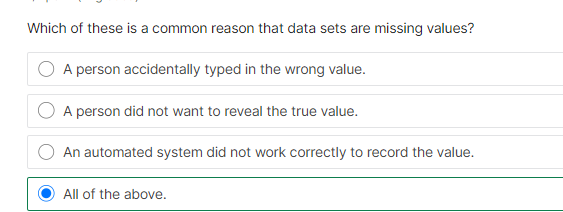
**example**1: for organ transplat, gender is part of standard record, but if they had headache a week before is not easily obtainable and more likely to be missed with questionaning

**Example**2: people with high income more likely to not mention the income

**Example3:** radargun detecting speed of the car is more likely to capture normal speed thatn a lower speed car



* BIAS IN MISSING DATA
* Example : heart transplant record will have date of dealth. When empty , shows successful transplantation

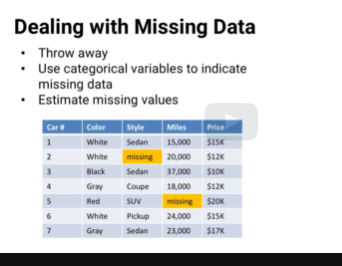


# 10.2 methods that donot require Imputation

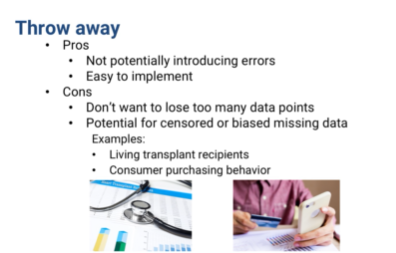
**Agenda**: ways to deal with missing data

3 ways

1. throw the data point
2. use categorical value
3. estimate missing value

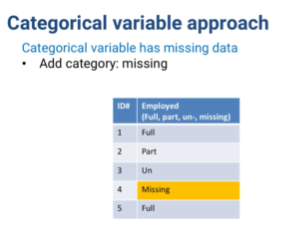


## Throw away



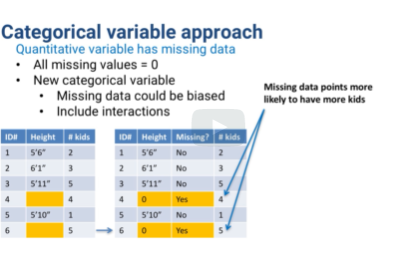
**Suggestion: check for patterns before deciding to throw the data out**

## use categorical value

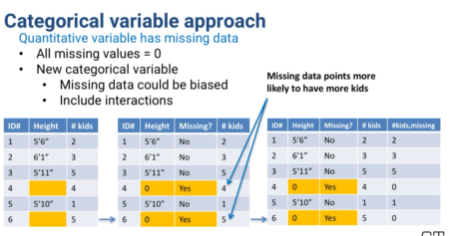


### For quantitative variable

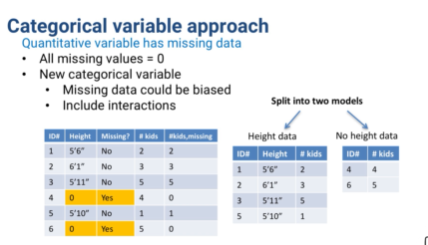
**Option#1:**

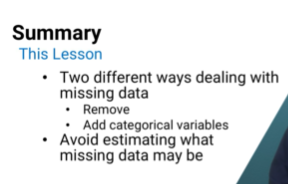


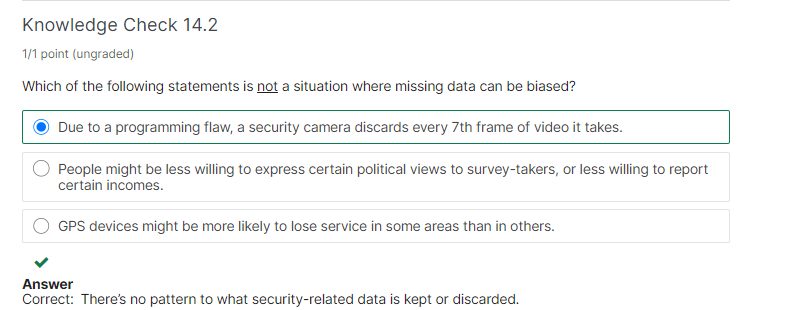
1. replace 0 and add categorical value showing missing flag
2. disadv: if sometypes of data is more likely to have missing data, then co-efficients of other variables will be pulled into one direction(meaning, the number of kids with missing data points has more kids)
3. Other Approach is to use Interaction variables (the 3rd table in the screenshot)



1. But, now we have two variables, # of kids when height is missing and # of kids when height is not missing.
2. Since we introuduced new categorical value, interaction terms with the new categorical variable and the existing variables will have two models (one with missing data in the set and other when there isn’t)
3. Its like a tree model with just one branch (shown below)



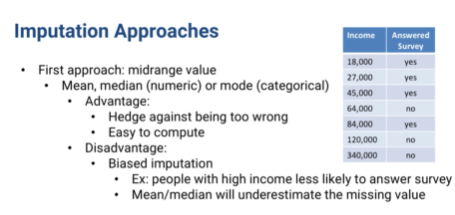




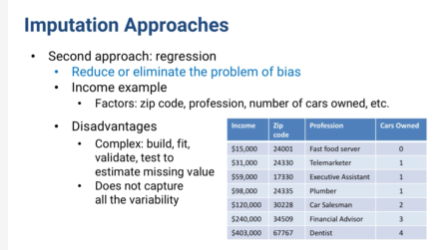
# 10.3 Imputation

**Agenda:** 3 ways to impute data (tradeoff simplicity & realism)

## Option#1: mean/median or mode for categorical



## Option#2:predict missing value using regression



**Adv:** it gives better value for missing data

**Disadv**:

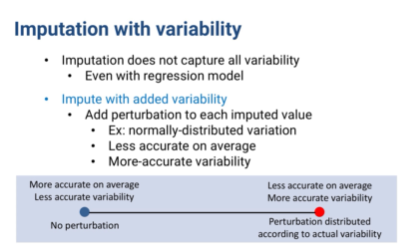
1. the data is being used twice, once for imputation and other for model fit. So we might overfit
2. doesn’t capture all variability. Example: Doctors in zip code, with 3 cars will have wide range of income.but predictive model will assign same income for all of them.

To handle variabililty, we can use **perturbation**

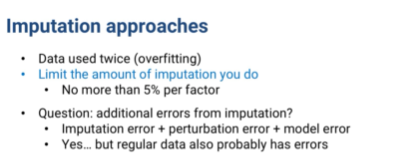
**Perturbation:** A random amount up or down from each imputed estimate.

**Example**: draw a normal distribution based on the error in the predictive model and change the imputed estimate by that amount

**Will adding perturbation make imputed value less accurate?** YES. But gives better estimate of the overall spread of the data



**Summary thus far:**

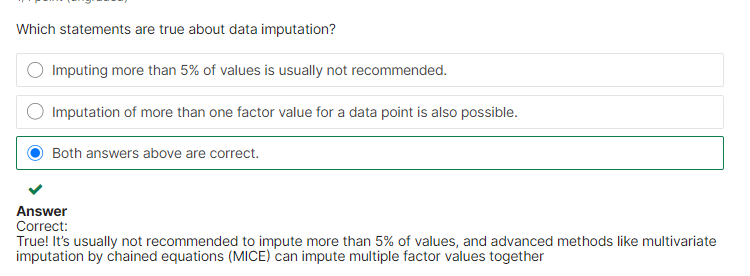
****

-we create a model with natural erros. Now we add more by perturbation and imputation. How to handle these compounding level of errors?

-Technical : not a way

**Suggestion :** can use test data. But not perfect if test data is missing value

\*WORRY as much as about the models with no missing data, because missing data is very common



# 10.4 Introduction to Optimization

Models that can be solved using Optimization

* Clustering, regression, variable selection, time series analysis and more.

**Usage:**

* descriptive and predictive analytics => ML models involve optimization to find their answers.
* Important for prescriptive analytics (give what I know and predict, what is the best course of action)

## Examples of questions that Optimization can answer

1) which airplane mechanic should be scheduled for each shift over the next week?

To meet the expected maintenance requirements at O'Hare Airportat lowest cost.

2) Making sure we don't violate any federal regulations or union contract requirements and accounting for unexpected breakdowns.

3) How much crude oil should be send by tanker and by pipeline from each oil field to each refinery to meet demand and avoid oversupply while keeping costs low?

4) What webpages should be optimized in a server farm? And how many copies of each should be stored to maximize the profit made by responding quickly to hits?

5) How should a large machine shop sequence its production to get maximum throughput while meeting all clients' deadlines?

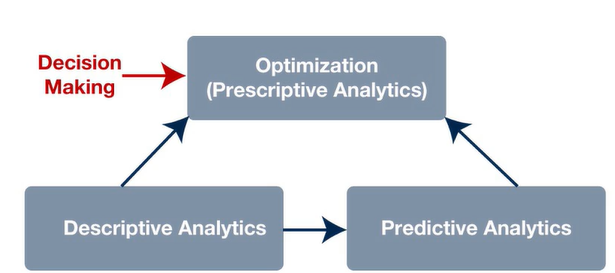
Taking into account the reality that some output will fail inspection and need to be remained.

6) even the GPS routing problem, what's the shortest route from my house to the airport, given current and predicted traffic?

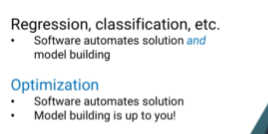


**About Optimization**

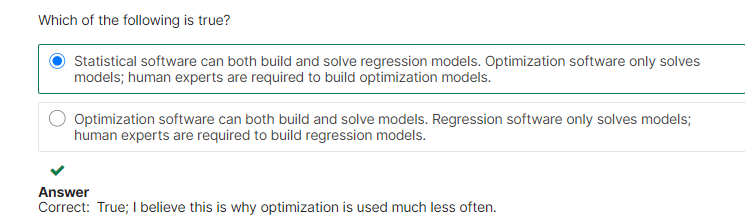
1. Powerful tool



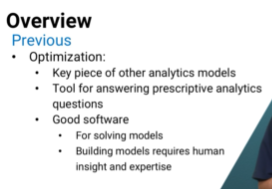
1. direct your organization at strategic, operational and tactical levels.
2. Difficult tool –meaning many software can automate solution to build and solve models. But Optimization model cannot be build automatically by a software..they can solve it







# 10.5 Elements of Optimization models



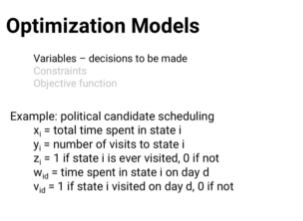
3 elements of optimization models



## 1.Variables

\* decision that optimization solver will pick the best value for

**Example:** travel plan of a political candidate

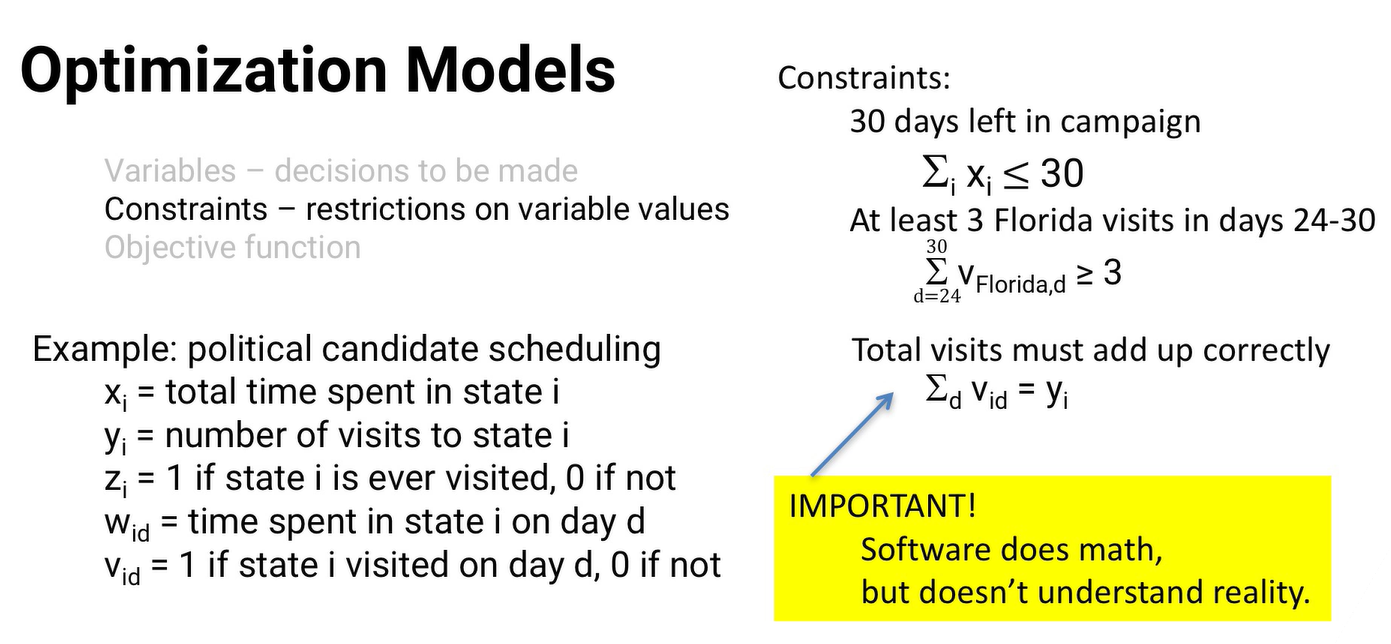


These are decisions we are trying to make

* if we build our model correctly, the optimization software will then find the best solution, the best combination of values for the variables

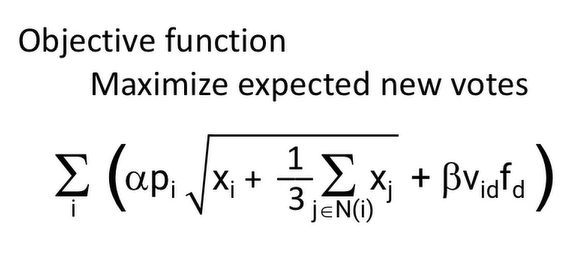
## 2. Constraints

\*restriction of the model



## 3. Objective Function

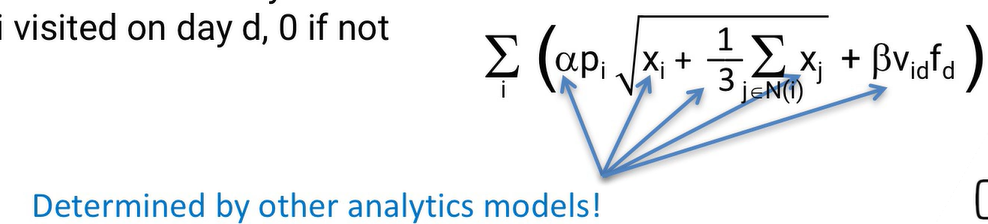
\*measure of quality of the model



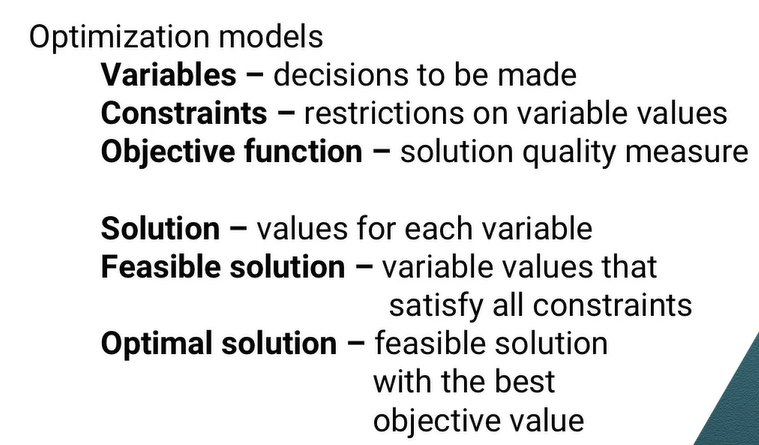
P(i) => population

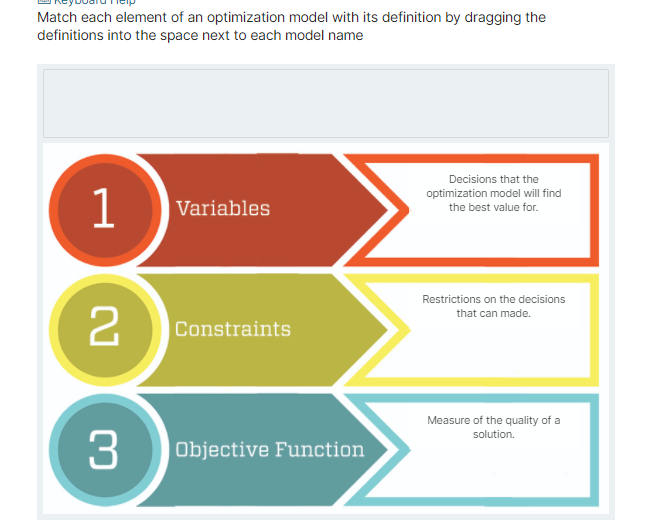
F => factor for the day(too early or too late visit)

* to use optimization, use output of other statistical model(like regression)



**WHAT is SOLUTION in Optimziation?** It is value for each variable



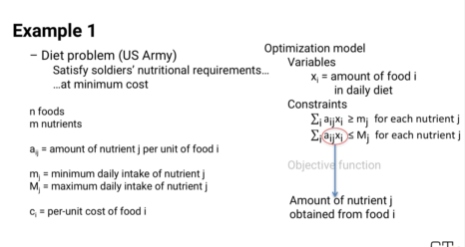


# 10.6 Modeling is an art

Agenda: examples of optimization model

## Example#1: Diet problem (Inituition can lead good models)

-1930 and 1940



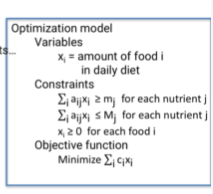
**Variables:** decision to make

X(i) =amount of food in daily diet (x(broccoli) is amount of broccoli)

**Constraint**

* For each nutrient “j”, a(ij) is amount of nutrient per food fr each nutrient\* x(i) is amount of food in daily diet
* impossible to each negative food , so x(i) >0

**Objective:** minmizse cost



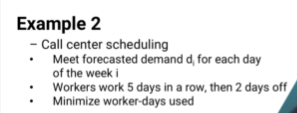
**Other variation**

1. cost variation : plums more abundant in summer and less expensive in summer
2. variety of food
3. combination of taste of food. PBJ!



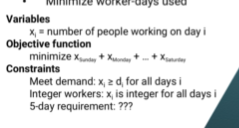
## Example2 : Intuition leads the wrong way

**Use case:**

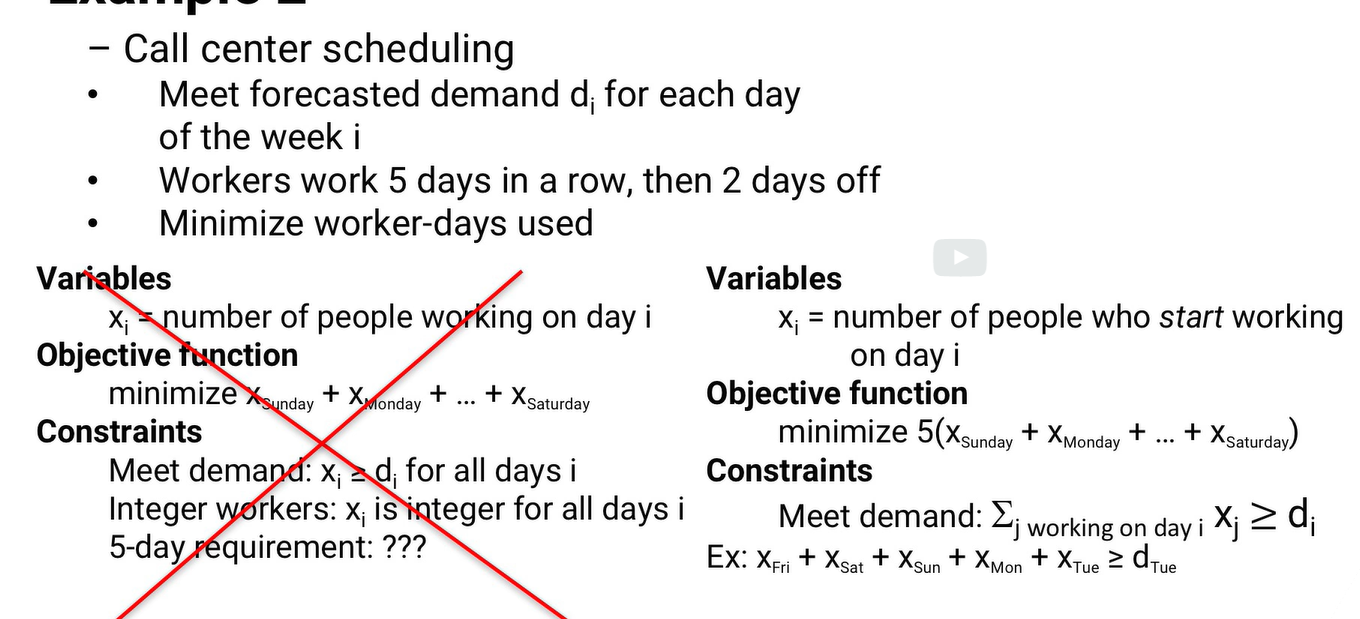


### Wrong:

1. Makes objective function easier: sum of 7 x)i) for 7 days is what we need to minimize
2. Constraints : meet demand



### Correct way:



**Variable:**

1. xi is the number of people who start their 5 working days in a row on day I.

**example:** x (Sunday)is the number of people who work Sunday, Monday, Tuesday, Wednesday, and Thursday, and then get Friday and Saturday off.

**Objective:** Each variable now accounts for 5 worker-days, so the total worker days is 5 times the sum of the seven xis.

(meaning, x(i) gives just the start day of a worker when he works 5 days. worker in total works for 5 days.so 5\*x(i) for all days will give the total work for workers which was what we want to minimize)

**Constraints:**

**Example:**

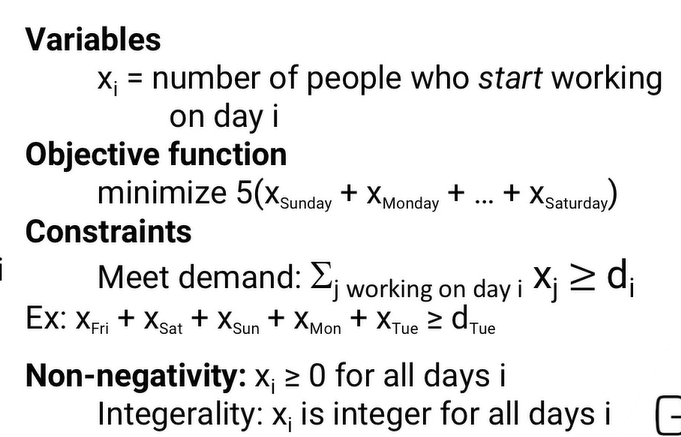
People who start their 5 days on Friday will still be working on Tuesday, as will people whose 5 days start on Saturday, Sunday, Monday, and of course, Tuesday.

So x Friday + x Saturday + x Sunday + x Monday + x Tuesday >= demand for Tuesday

So, summarize it, to find people working on each day, we can add 5 day “x” value

**Two more constaints**

1. cannot hire fraction of worker (so x(i) is integer
2. for all days, the number of people working should be greater than zero



# 10.7 Modeling with Binary variables

**Agenda:**

In this lesson we are going to see how to create more complex structures in optimization models.

By introducing binary variables to the new model.

**What is binary variable to Optimization model**

**Objective**: Given a budget of B dollars, a person wants to invest in a way that balances expected return with investment risk.

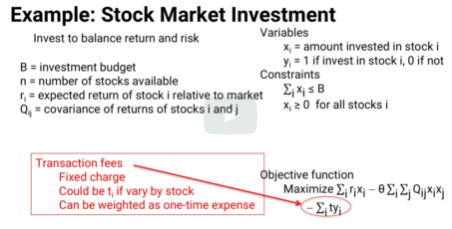


1. Constraint: short selling is not allowed ; so constraints x(i) >=0
2. Objective: we have binary theta times risk factor..by varying theta, we can find different solutions, different investments that balance return and risk differently

This sort of analysis helped H**enry Markowitz win the Nobel Prize in Economics.**

**For practical purpose:**

* Need to pay a transaction fee per stock that we invest in (It might not matter how much we buy, just whether we buy it or not.)
* Variables only tell how much we bought , not whether we bought it or not. So new variable y(i) to show binary response **(Binary variable)**
* We can create “t” (transaction fees) if we buy stock(if y(i)=1) to incorporate this function in the objective. This is called **FIXED CHARGE**



* **Linking Constraints with other variables**
* 1. If we invest in stock “I”, we have to pay the **transaction** cost.



If “y(i)=0” , then we cannot invest in stock at all

If “y(i)=1” , then we CAN invest upto budget “B” in the stock

* 2. We might also want to have some minimum amount that we invest in each stock. (We cannot invest 3 cent in Google as stock price is more)



* M(i)=> some minimum dollar amount for each stock

If y(i) =0, then we don’t invest at all in stock, which means no x(i)

If y(i) =1, then we need to atleast invest mimnimum stock proce in that stock

* Personal constraint: atleast one investment in testla



* Either Amazon , google or apply.if we want exactly one, then we can use =1 instead of >= 1



* Both fed and UPS or (neither fedex or upd), use below



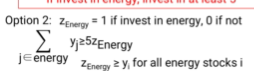
* Opposite investment. Either cola or pepsi



* Complex: if we invest in a energy stock, invest in atleast 3



Option#2: create new binary variable



To make sure that Z Energy really is 1, if we invest in an energy stock.

So for each energy stock i, the energy must be greater or equal to yi.

