PA 446

Coding for Civic Data Applications

Will be starting at 6:05pm

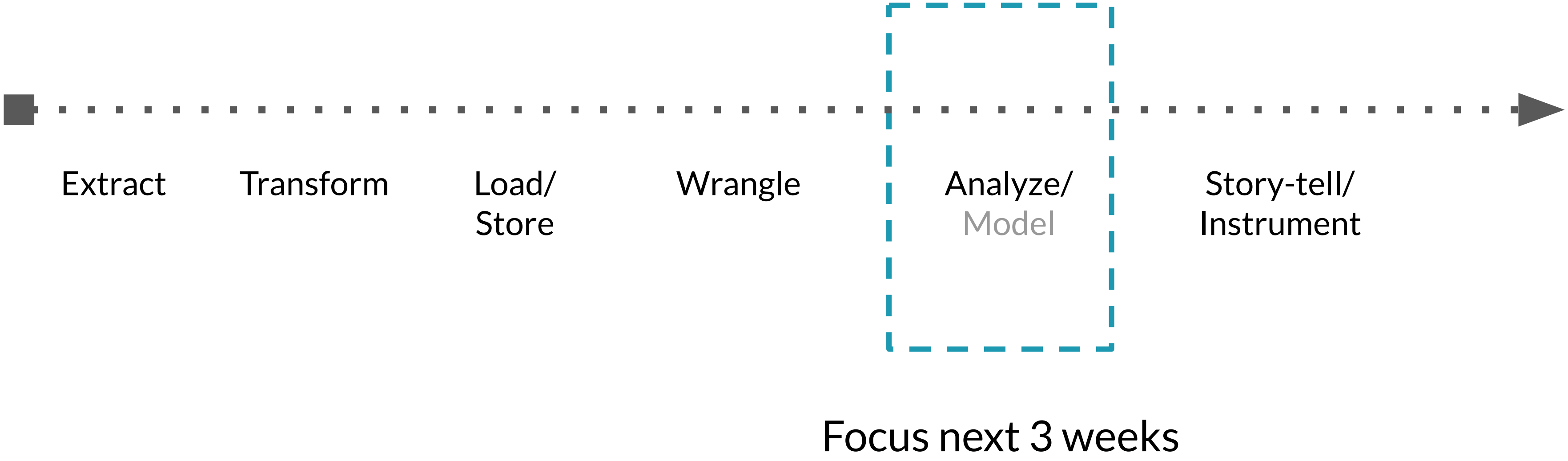
Class #6

Logistics

Course Logistics

- HW 4
  - Due 10/20
  - Available end of week
- HW 3, graded by EOW
- Midterm
  - Similar format as HW4
  - With a tighter time limit

Data Science “workflow”



Where We Been: Gender Data

1. Two variable comparison
2. Two variable significance testing

What We Found: Gender Data

Dept	**Difference Between Female and Male Salaries**	Female Salary Averages	Male Salary Averages
POLICE	**-4834.963677**	87338.2753	92173.239
FIRE	**-2537.510887**	103619.36	106156.871
STREETS & SAN	**-7476.160747**	68309.8452	75786.006
WATER MGMNT	**-11322.18714**	86732.6581	98054.8453
AVIATION	**-9040.943804**	73139.6085	82180.5523

How Confident Were We: Gender Data

Dicuss

How Confident Were We: Gender Data

Dept	**Difference Between Female and Male Salaries**	Female Salary Averages	Male Salary Averages	**p.value**	alternative
POLICE	**-4834.963677**	87338.2753	92173.239	**1.96E-38**	two.sided
FIRE	**-2537.510887**	103619.36	106156.871	**0.03008956**	two.sided
STREETS & SAN	**-7476.160747**	68309.8452	75786.006	**1.54E-08**	two.sided
WATER MGMNT	**-11322.18714**	86732.6581	98054.8453	**2.92E-24**	two.sided
AVIATION	**-9040.943804**	73139.6085	82180.5523	**3.02E-11**	two.sided

Takeaways: Gender Data

So What	Next Steps

Where We Been: Race Data

1. Multiple variable comparison
2. Multiple variable significance testing

Where We Been: Race Data

Department	api	black	hispanic	white	NA
AVIATION	89023	63501	76966	82891	77932
FIRE	90779	99732	99272	107545	106035
POLICE	85783	90703	86518	93484	90574
STREETS & SAN	66412	68870	72859	74037	72828
WATER MGMNT	94739	92964	93684	97413	96131

How Confident Were We: Race Data

Dicuss

Where We Been: Race Data

[to coding]

How Confident Were We: Race Data

Department	api	black	hispanic	white	NA	p.value
AVIATION	89023	63501	76966	82891	77932	5.19E-06
FIRE	90779	99732	99272	107545	106035	1.58E-11
POLICE	85783	90703	86518	93484	90574	3.37E-52
STREETS & SAN	66412	68870	72859	74037	72828	0.029837002
WATER MGMNT	94739	92964	93684	97413	96131	0.000458519

How Confident Were We: Race Data

Department	api	black	hispanic	white	NA	p.value
AVIATION	89023	63501	76966	82891	77932	5.19E-06
FIRE	90779	99732	99272	107545	106035	1.58E-11
POLICE	85783	90703	86518	93484	90574	3.37E-52
STREETS & SAN	66412	68870	72859	74037	72828	0.029837002
WATER MGMNT	94739	92964	93684	97413	96131	0.000458519

Takeaways: Race Data

So What	Next Steps

Takeaways: Race Data

What Can We Do?

In Order to Say Anything About 1 Race Relative to Another

Department	api	black	hispanic	white	NA	p.value
AVIATION	89023	63501	76966	82891	77932	5.19E-06

Salary Averages

api
black
hispanic
white

Salary Averages

api
black
hispanic
white

In Order to Say Anything About 1 Race Relative to Another

Department	api	black	hispanic	white	NA	p.value
AVIATION	89023	63501	76966	82891	77932	5.19E-06

Salary Averages

api
black
hispanic
white

Salary Averages

api
black
hispanic
white

In Order to Say Anything About 1 Race Relative to Another

Department	api	black	hispanic	white	NA	p.value
AVIATION	89023	63501	76966	82891	77932	5.19E-06

Salary Averages

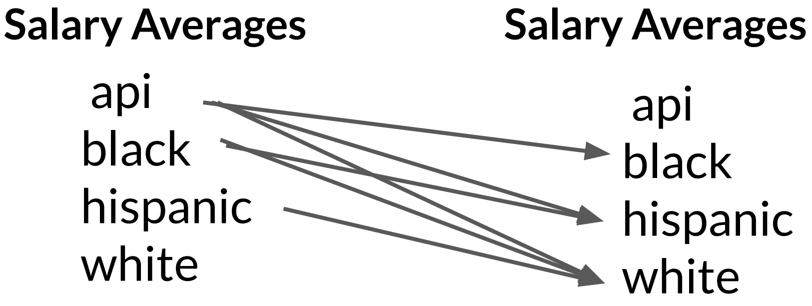
api
black
hispanic
white

Salary Averages

api
black
hispanic
white

In Order to Say Anything About 1 Race Relative to Another

Department	api	black	hispanic	white	NA	p.value
AVIATION	89023	63501	76966	82891	77932	5.19E-06



Number of tests : (3+2+1)* # of Departments = 30 tests
A lot of work + error prone

In Order to Say Anything About 1 Race Relative to Another

What Can We Do?

In Order to Say Anything About 1 Race Relative to Another

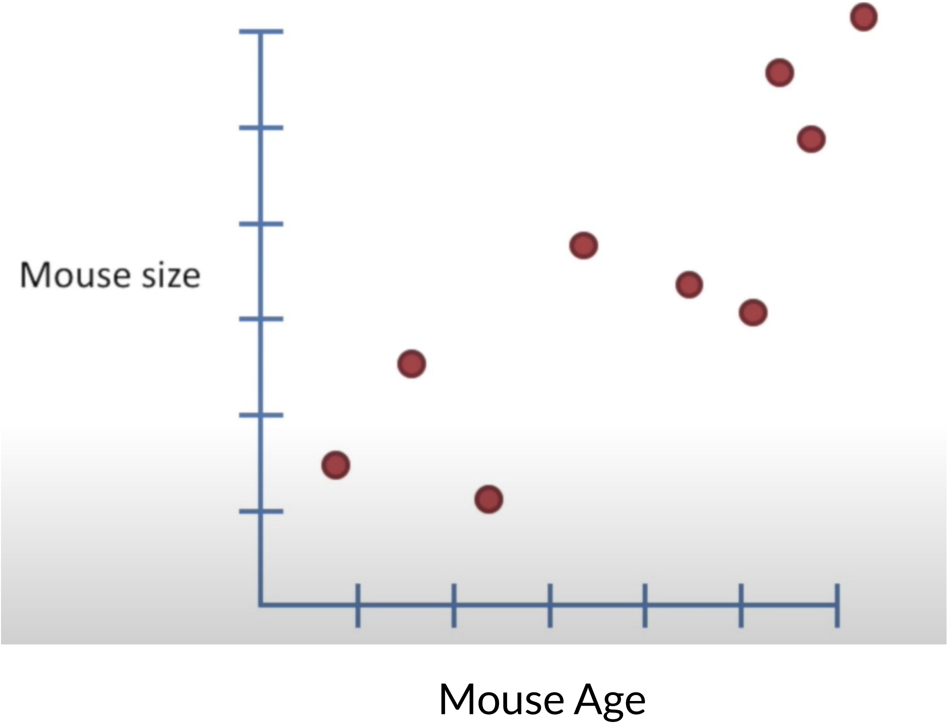
What Can We Do?

- Clarify scope with the client and compare less variables
- Use linear regressions

Linear Regression

Basic Linear Regression

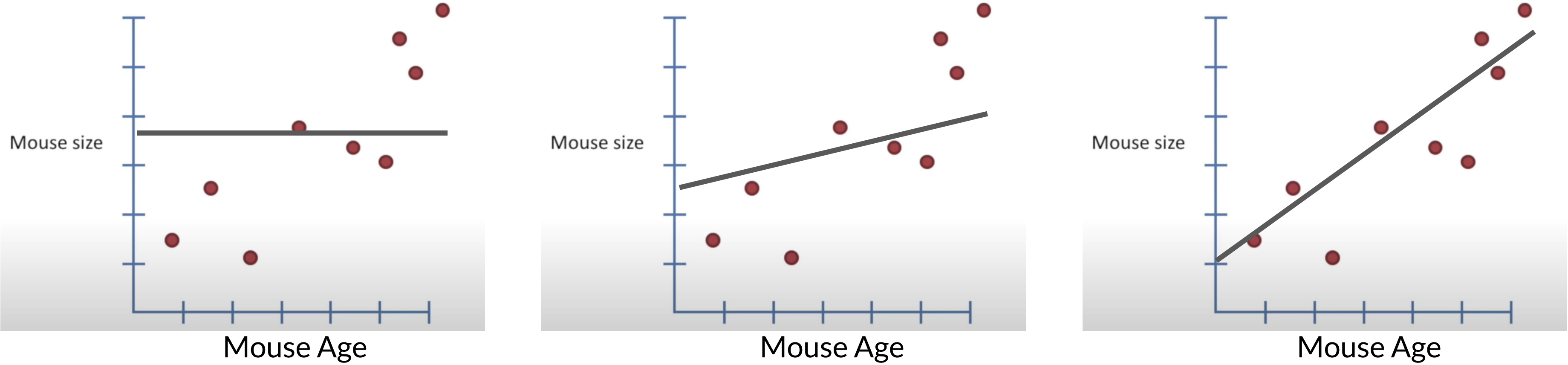
Purpose



How does change in Mouse age impacts Mouse Size?

Basic Linear Regression

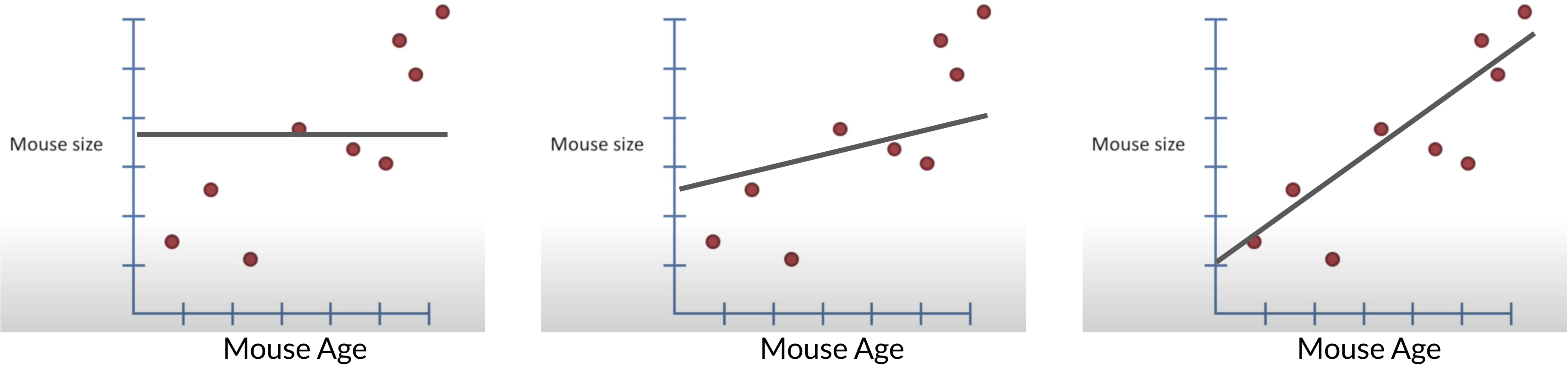
Finding a Line Best Fit



A lot of compute power is used testing out different lines and slopes until the line best fitted to our data is determined

Basic Linear Regression

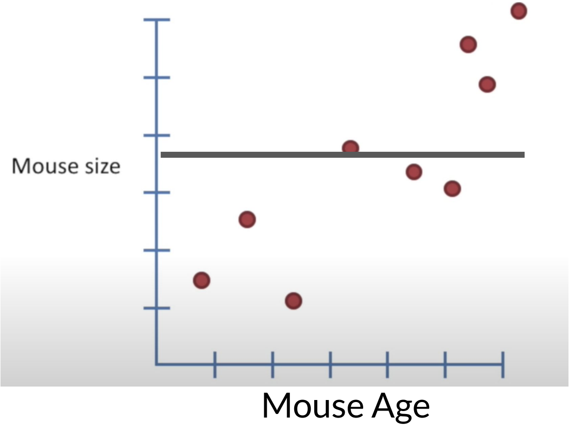
Described in Equation Form



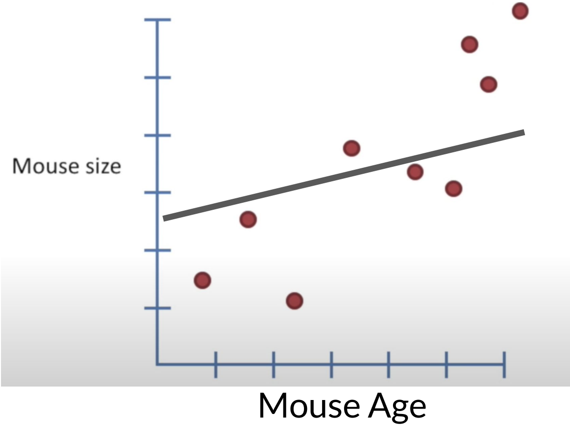
y = slope* X + constant (y-intercept)

Basic Linear Regression

Finding Relationships Between X and Y



y = 0* X + constant_a



y = 0.5* X + constant_b

y = 1.1* X + constant_c

y = Mouse Size X = Mouse Age Constant = y-intercept

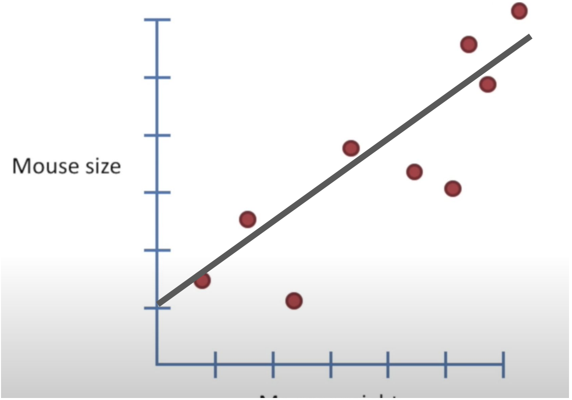
Basic Linear Regression

Interpreting the Formula

y = 1.1* X + constant_c

1.1* X: a 1 unit increase in mouse age is correlated with a 1.1 unit increase in mouse size

constant_c: at age 0, a mouse has size 1



y = 1.1* X + constant_c

New Question

Within the **Same Sample**, Another Variable Impacts Size

y = Mouse Size

X2 = Whether the mouse received gene therapy (Control vs Mutant)

Now what you want to know is **how gene therapy impacts the size of the mouse**

New Question

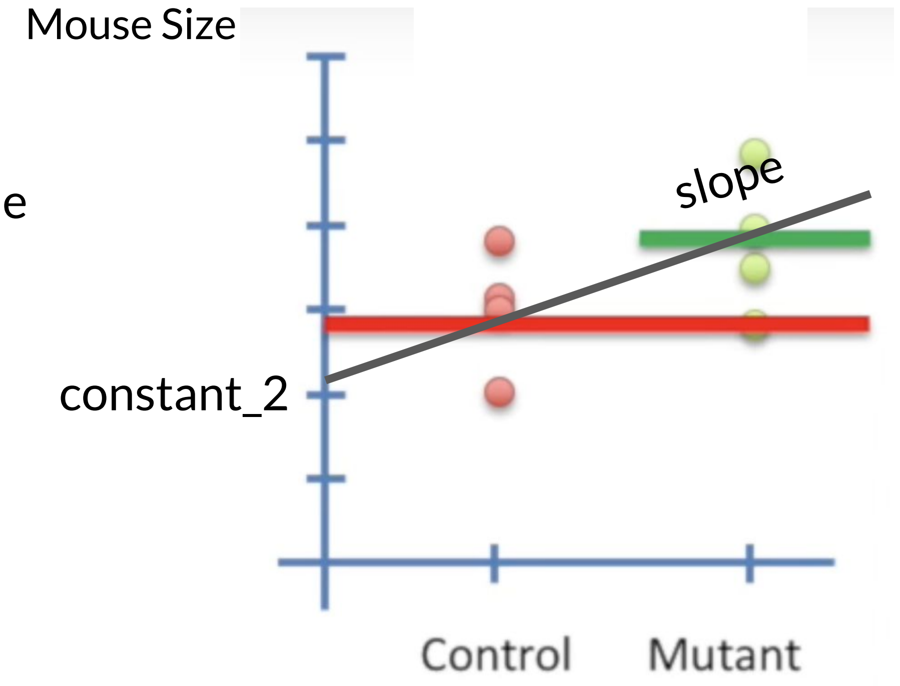
Within the **Same Sample**, Another Variable Impacts Size

y = Mouse Size

X2 = Whether the mouse received gene therapy

X2 is a categorical variable (Control or Mutant). How do you represent it in the linear regression equation?

Y = slope * X2 + constant_2



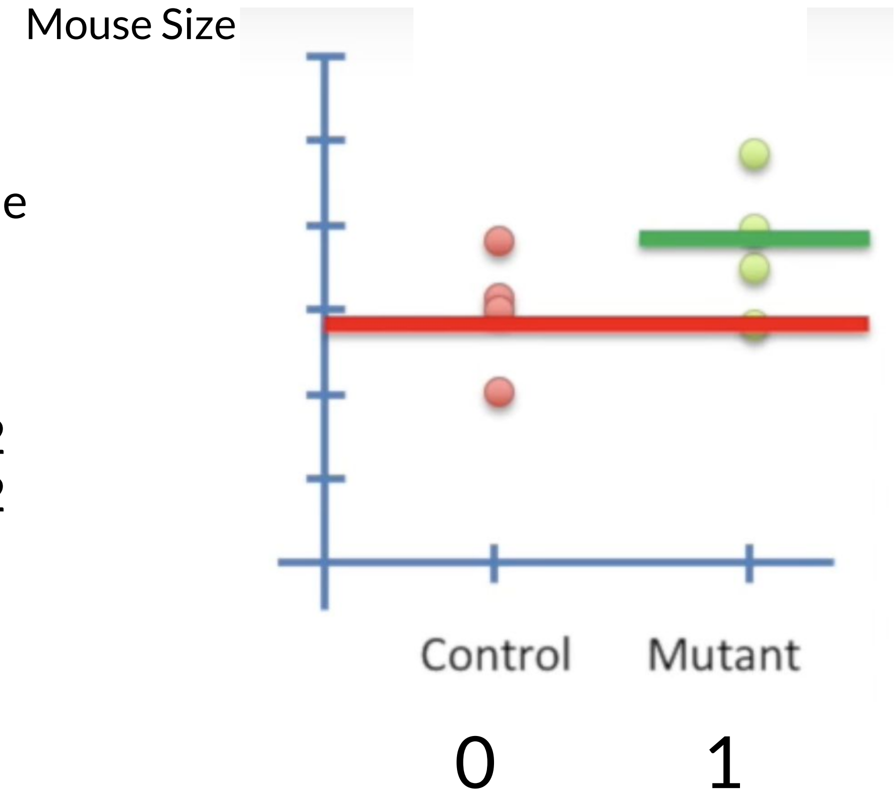
Categorical X Values Must be Converted to “Dummy” Vars

Control or Mutant → 1 or 0

y = Mouse Size

X2 = Whether the mouse received gene therapy

When X2=0, Y = slope * 0 + constant_2
When X2=1, Y = slope * 1 + constant_2



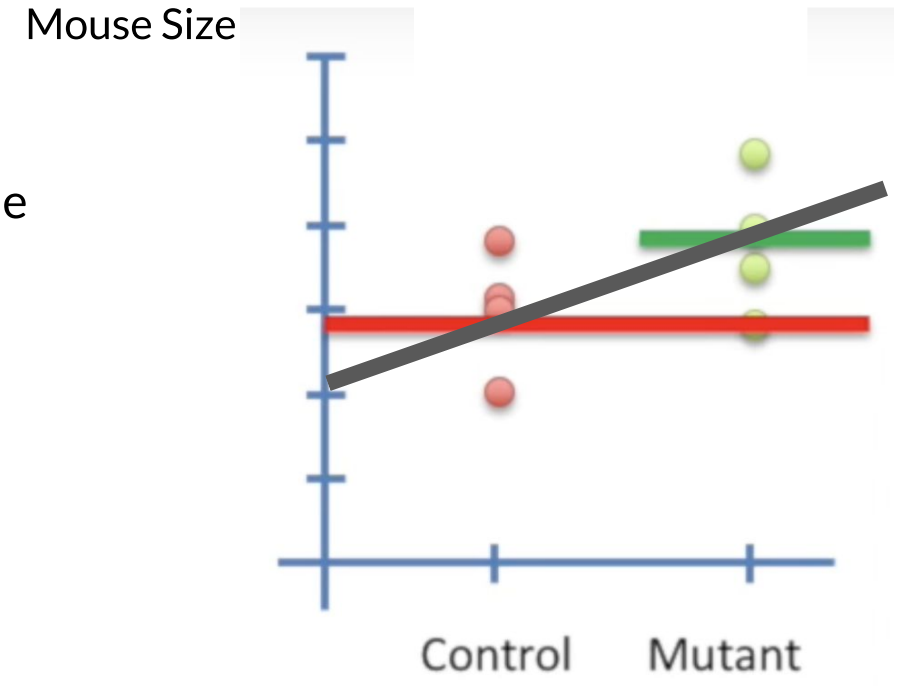
New Question

Another Linear Regression

y = Mouse Size

X2 = Whether the mouse received gene therapy

Y = 0.5* X2 + constant_2

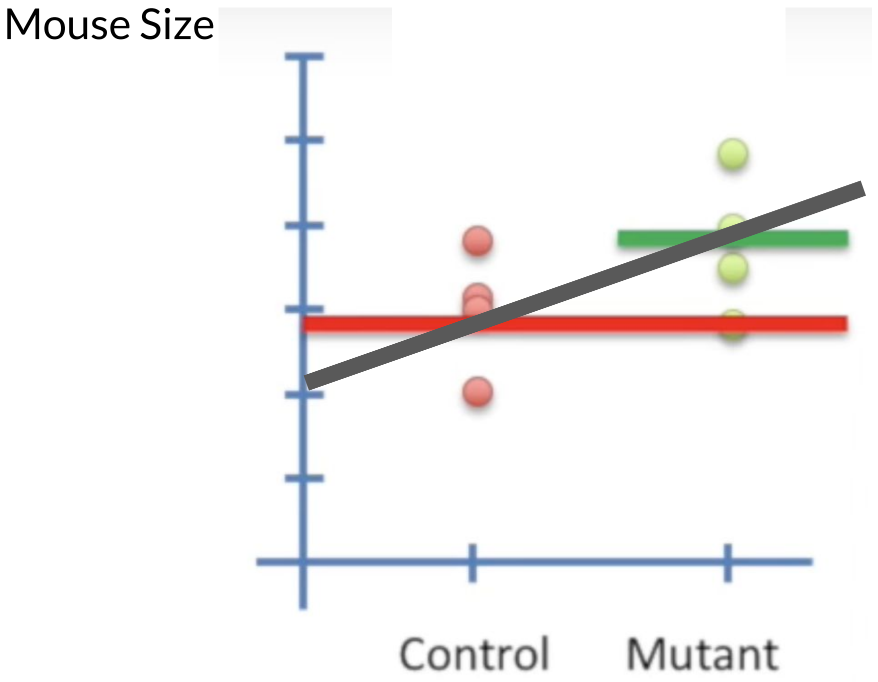


New Question

Complications

You just determined the relationship between gene therapy and mouse size (Y = 0.5* X2 + constant_2).

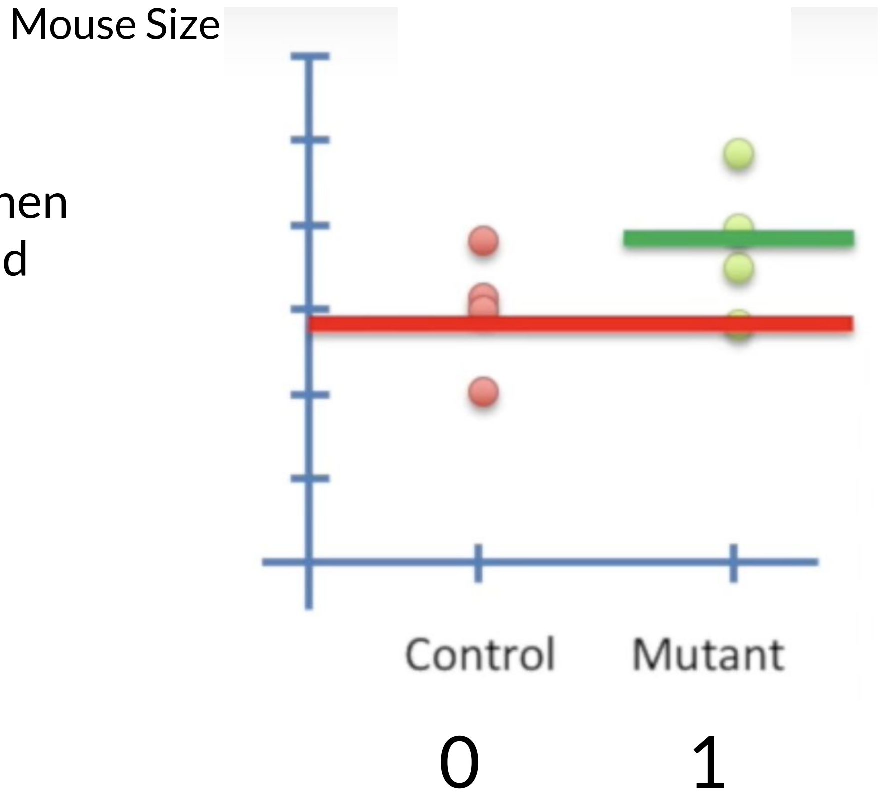
But remember, you are using the same sample as before. What might you have overlooked?



New Question

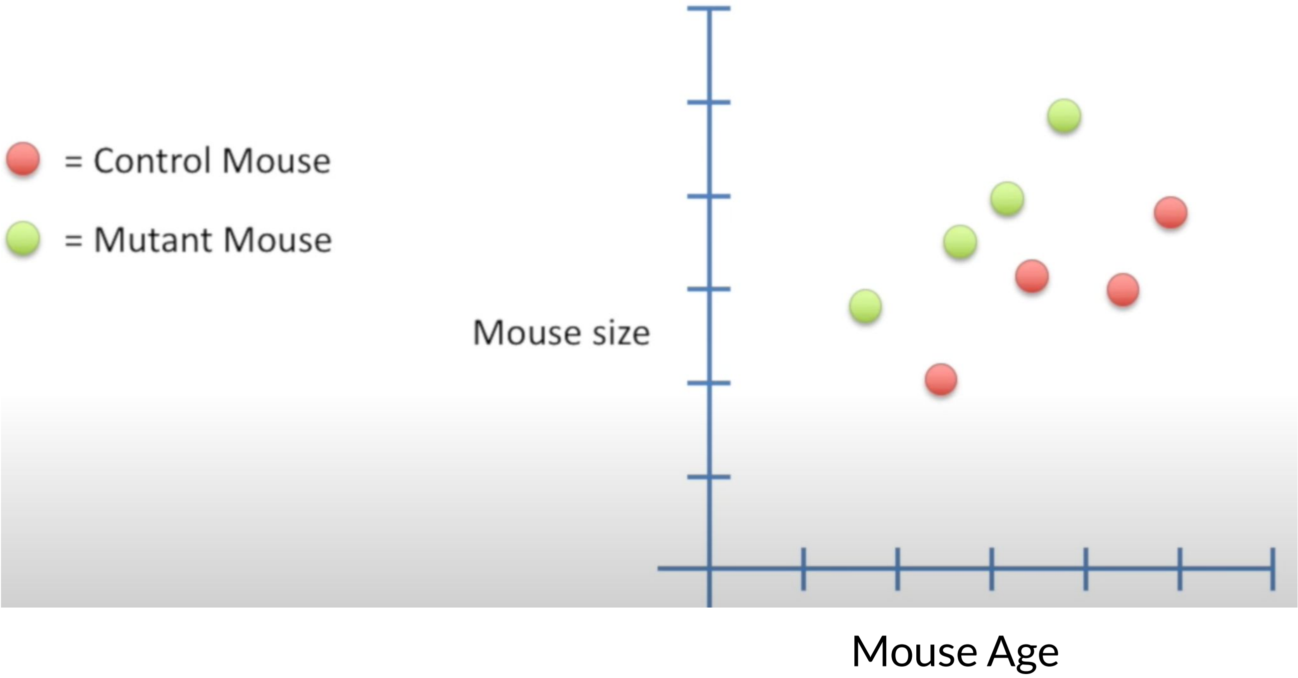
How to Control for Age?

Is there some way to control for age, when comparing the mouse size of control and mutant mice?



Multiple Regression

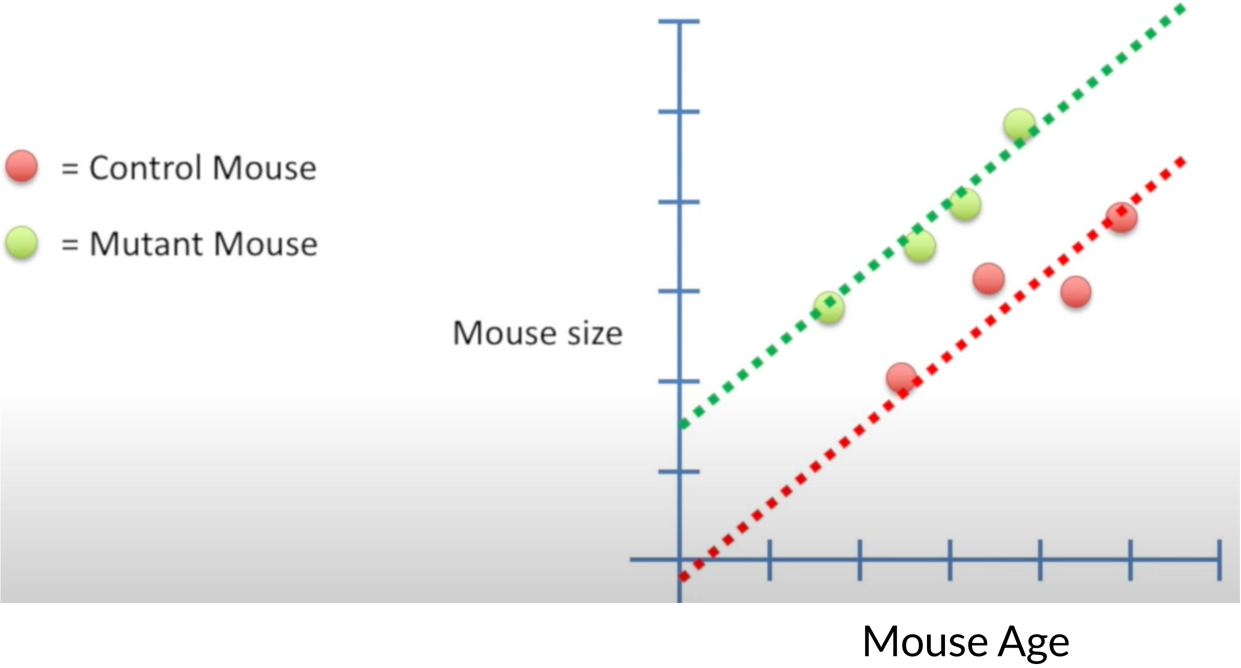
Finding Relationships Between X and Y, Controlling for X2



What relationship do you see?

Multiple Regression

Finding Relationships Between X1, X2 and Y



Formula

y = 1.1* Mouse age + **0.5*Mutant mouse** + constant

Interpretation:
Mutant mice are 0.5 size larger than control mouse, **controlling for mouse age**

A 1 unit increase in mouse age is correlated with a 1.1 unit increase in mouse size, **controlling for the mouse' gene status**

Multiple Regression

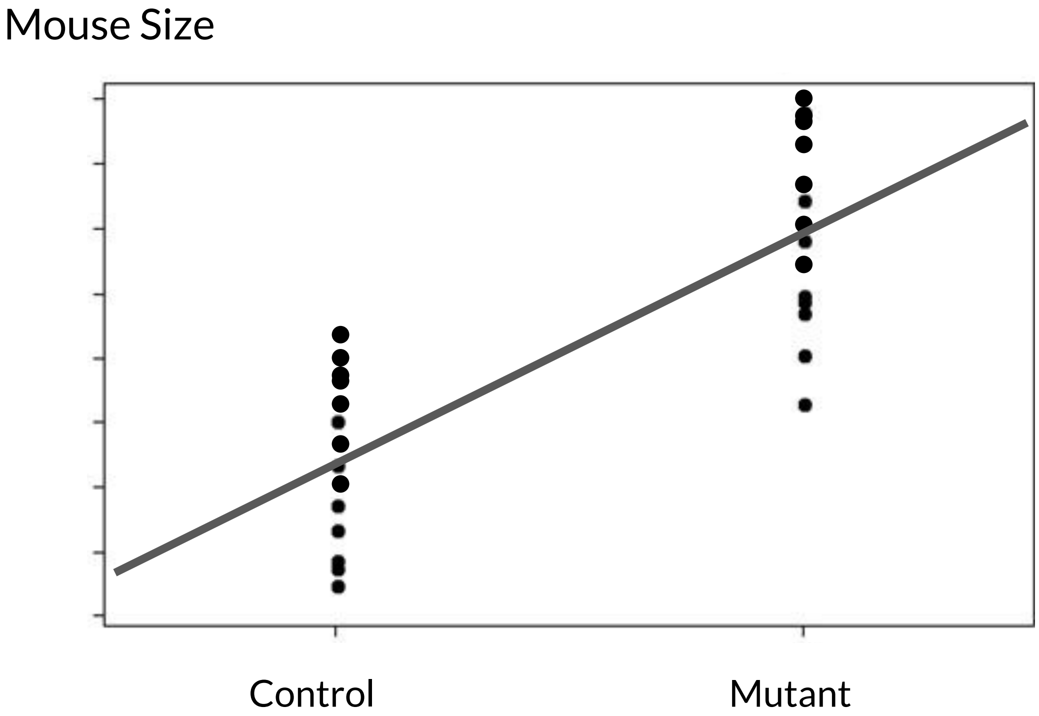
Finding Relationships Between X1, X2 and Y

Linear regressions works for numeric X's and/or categorical X's

Let's try 1 more time with exclusively categorical X's: mice gene therapy and gender

Multiple Regression: Categorical

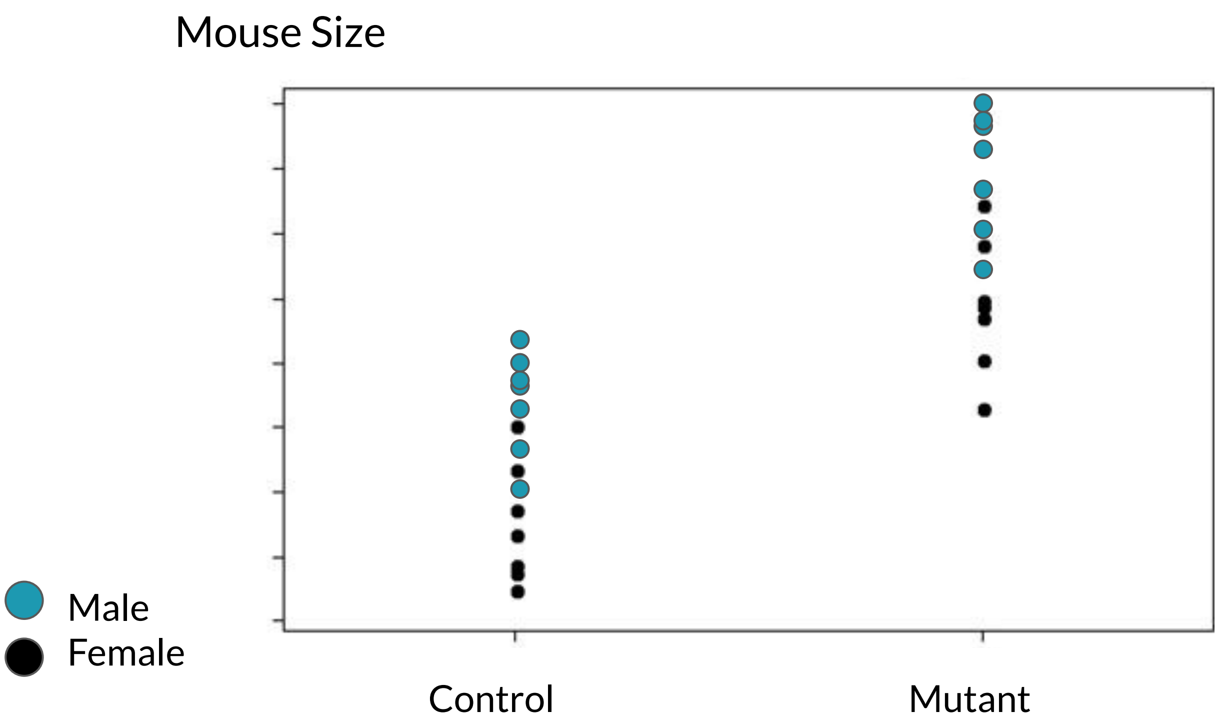
Multiple Binary Variables: Control v Mutant



y = **1.1*Mutant mouse** + constant

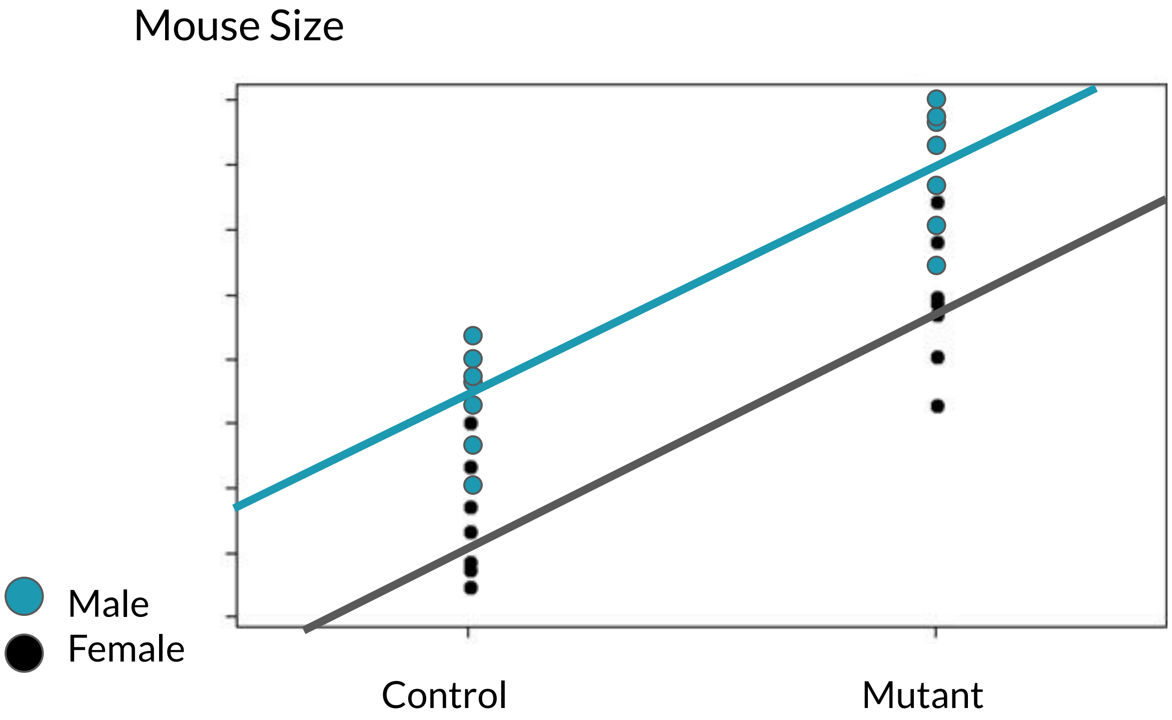
Multiple Regression: Categorical

Multiple Binary Variables: Control v Mutant and Male v Female



Multiple Regression

Multiple Binary Variables: Control v Mutant and Male v Female



Final Formula

y = 1.5*Mutant mouse
+ **2*gender**
+ constant

Interpretation:
Male mice are 2 sizes larger than female mice, controlling for gene status

Mutant mice are 0.5 sizes larger than control mice, controlling for gender

10-minute break, be back by 7:15pm

How Does This Help with Our Problem with Race?

Discuss

Mouse Size	Gender (1= M, 0=F)	gene type (1=Mutant, 0 = Control)	mouse age
4	1	1	2.4
3.6	1	0	1.5
6	0	0	3.0
4.8	0	1	3.8
5	1	1	5.1
3	0	0	2.2

Salary	Department	race	gender
118998	POLICE	white	M
109662	FIRE	black	M
121272	DAIS	NA	NA
119712	WATER MGMNT	API	M
92352	TRANSPORT	hispanic	M
72510	POLICE	NA	F

I omitted name and job title

Wrangling Needed Before Running a Regression

Discuss

1.

Generating Dummy Variables

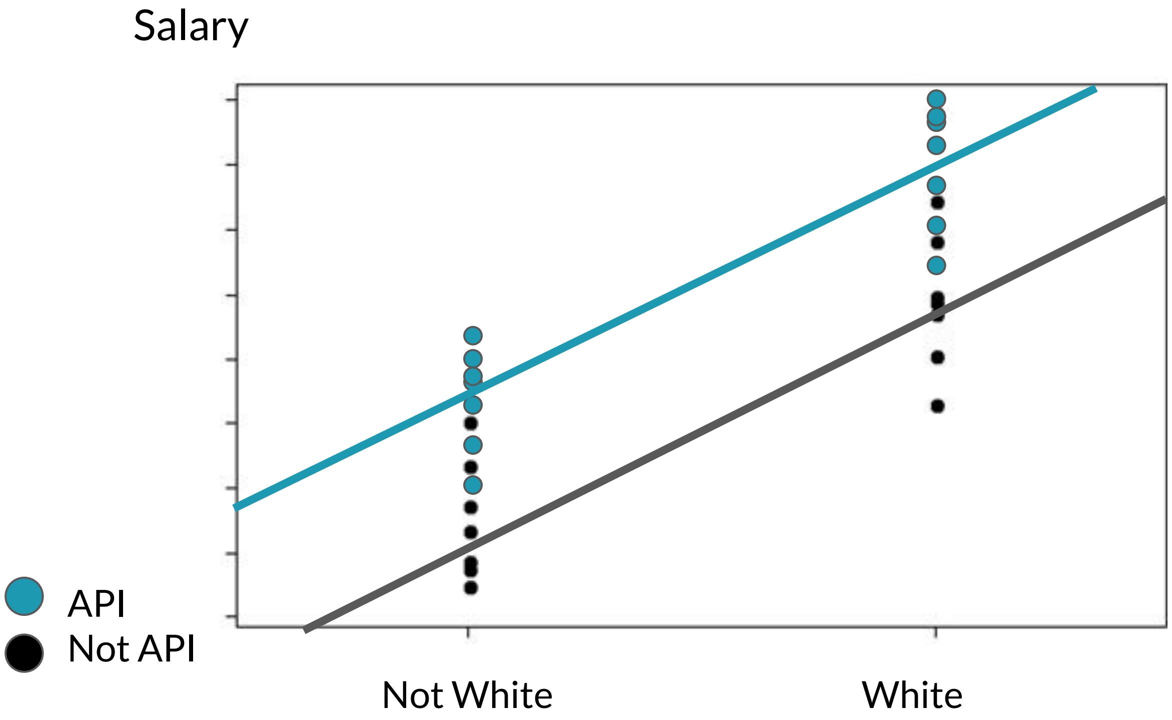
Discuss

race
white
black
white
API
hispanic
NA

race_white	race_black	race_api	race_hispanic	race_aian
1	0	0	0	0
0	1	0	0	0
1	0	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	0

Multiple Regression

What If We Isolated to Only White and API



Final Formula

Much harder to visualize multiple regressions after 2 variables because you will need a 3D space

Data Quality Checks for Linear Regressions

NA's Are Usually Bad

1. Impute (average, median etc)
2. Remove the NA rows
3. Remove the columns with NAs

Data Quality Checks for Linear Regressions

No categorical variables for linear regression

All numerics

- Dummy (1's and 0's) - remember to exclude 1 category, otherwise you have collinearity
- Doubles (any numbers)

Applying to Race Data

Data Quality Checks for Linear Regressions

No categorical variables for linear regression

All numerics

- Dummy (1’s and 0’s) - remember to exclude 1 category, otherwise you have collinearity
- Doubles (any numbers)

Regression Results: Aviation Department

term	estimate	std.error	statistic	p.value
(Intercept)	83247.2	1064.9	78.2	0.00
final_race_two_api	6283.4	4665.3	1.3	0.18
final_race_two_black	-12266.4	2799.8	-4.4	0.00
final_race_two_hispanic	-6127.9	1692.4	-3.6	0.00

Interpretation:

Regression Results: Aviation Department

term	estimate	std.error	statistic	p.value
(Intercept)	83247.2	1064.9	78.2	0.00
final_race_two_api	6283.4	4665.3	1.3	0.18
final_race_two_black	-12266.4	2799.8	-4.4	0.00
final_race_two_hispanic	-6127.9	1692.4	-3.6	0.00

Regression Results: Aviation Department

term	estimate	std.error	statistic	p.value
(Intercept)	83247.2	1064.9	78.2	0.00
final_race_two_api	6283.4	4665.3	1.3	0.18
final_race_two_black	-12266.4	2799.8	-4.4	0.00
final_race_two_hispanic	-6127.9	1692.4	-3.6	0.00

Interpretation: the average salary of final_race_two_white = 1

Regression Results

term	estimate	std.error	statistic	p.value
(Intercept)	83247.2	1064.9	78.2	0.00
final_race_two_api	6283.4	4665.3	1.3	0.18
final_race_two_black	-12266.4	2799.8	-4.4	0.00
final_race_two_hispanic	-6127.9	1692.4	-3.6	0.00

Interpretation: the average salary of final_race_two_api = 1 is $6283.4 higher than final_race_two_white = 1. The result is not statistically significant at 95% confidence level

Data Quality Checks for Linear Regressions

No categorical variables for linear regression

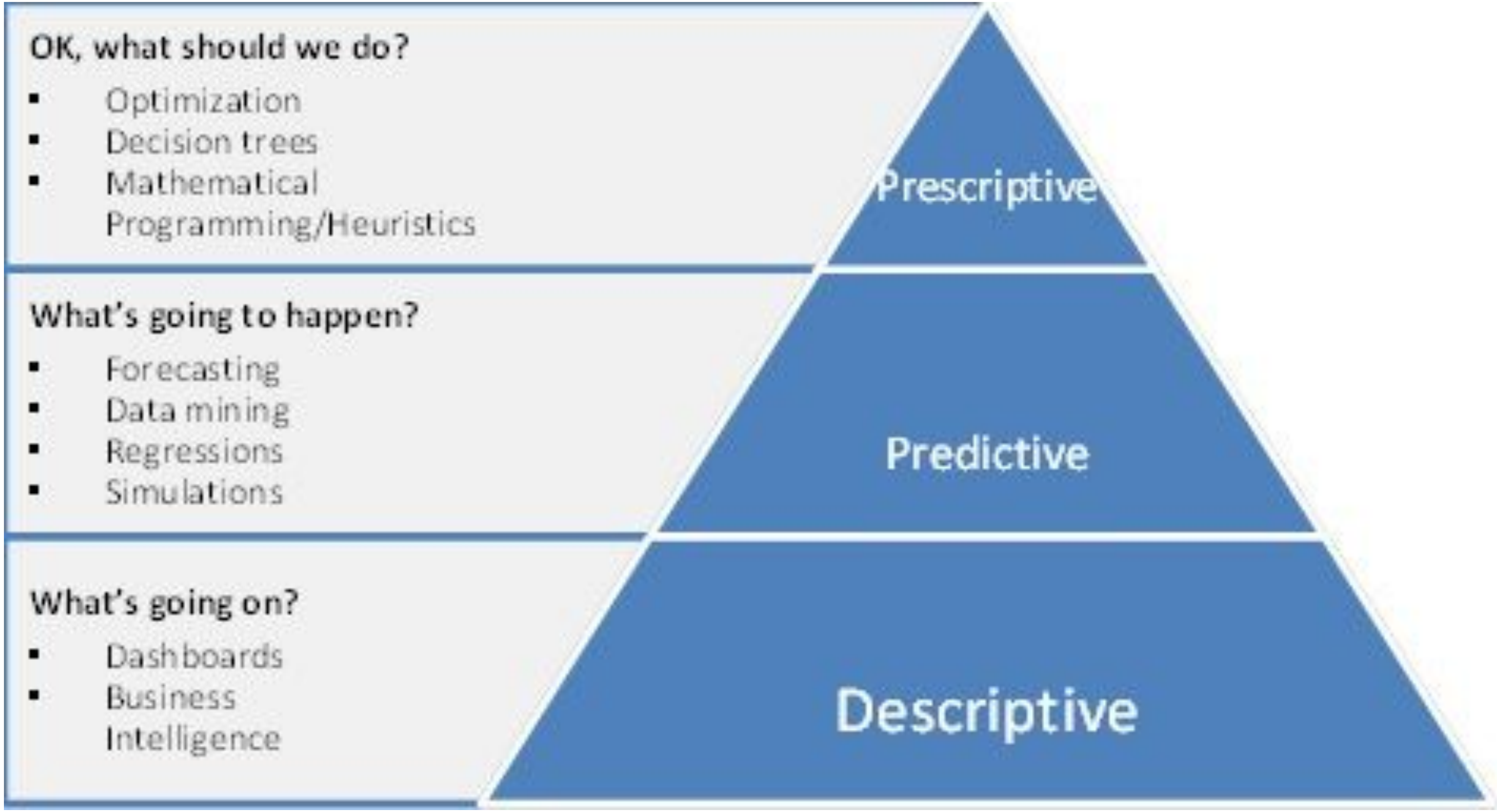
term	estimate	std.error	statistic	p.value
(Intercept)	83247.2	1064.9	78.2	0.00
final_race_two_api	6283.4	4665.3	1.3	0.18
final_race_two_black	-12266.4	2799.8	-4.4	0.00
final_race_two_hispanic	-6127.9	1692.4	-3.6	0.00

Descriptive

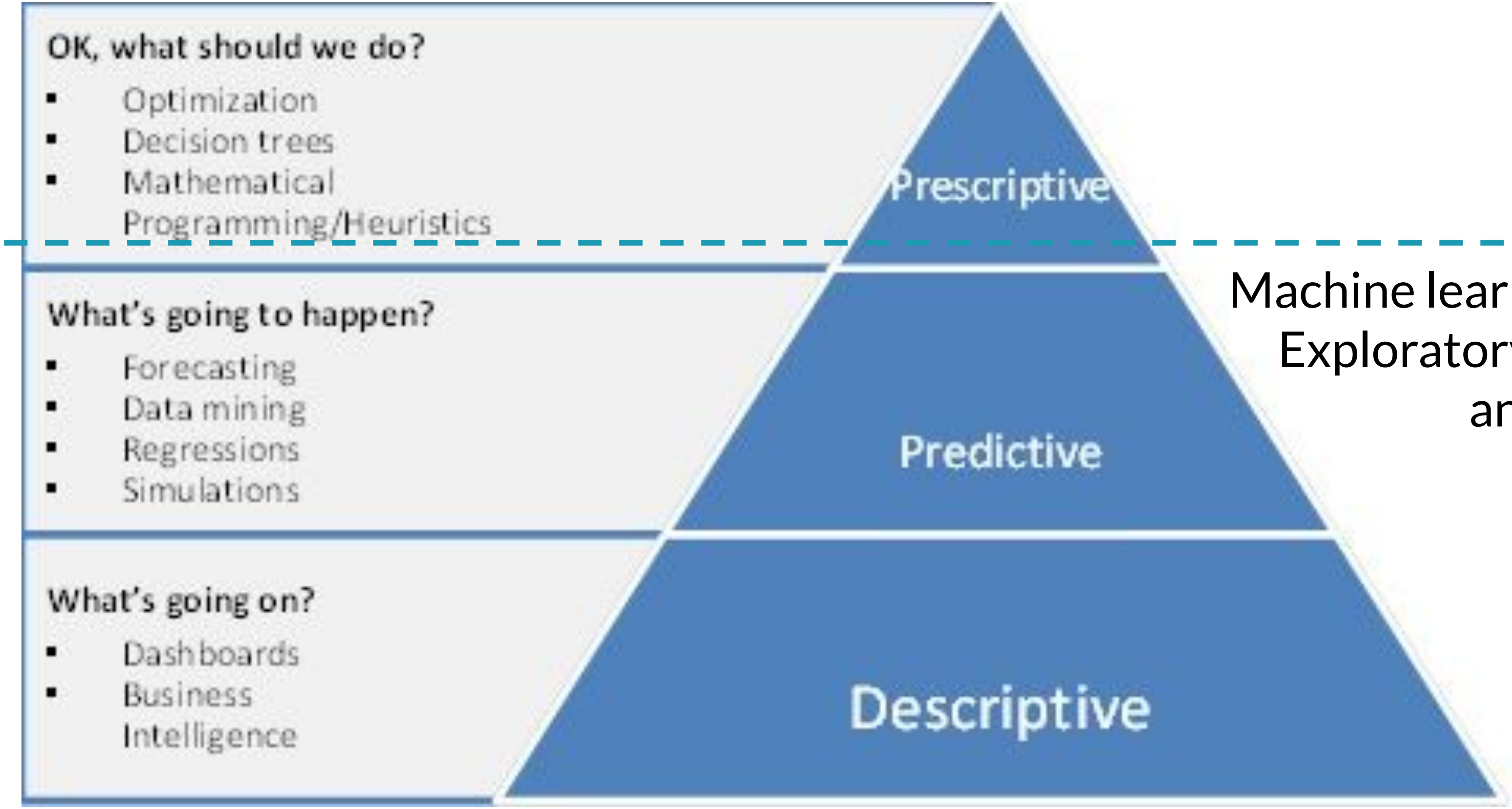
Predictive

Prescriptive

Hierarchy of Analysis

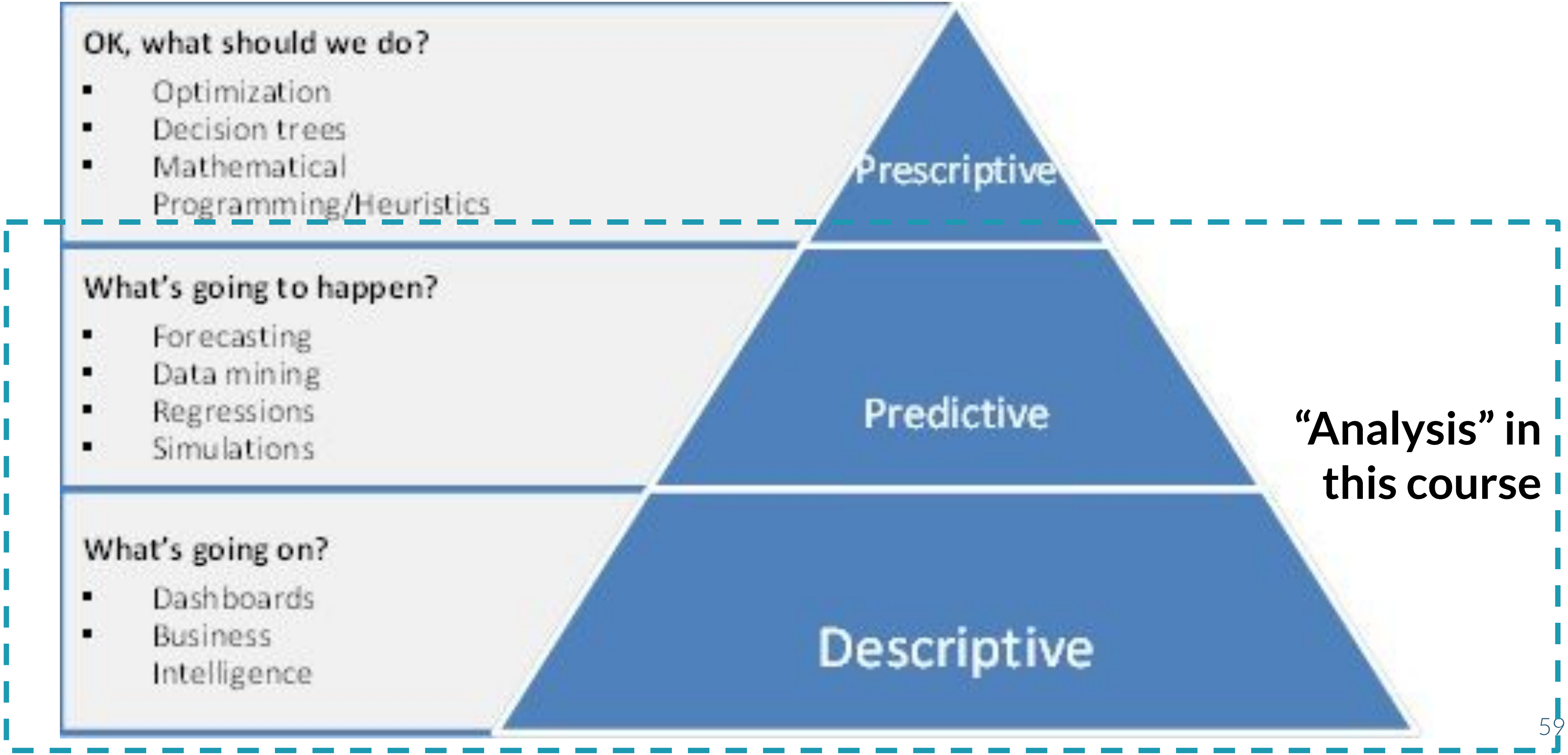


Hierarchy of Analysis

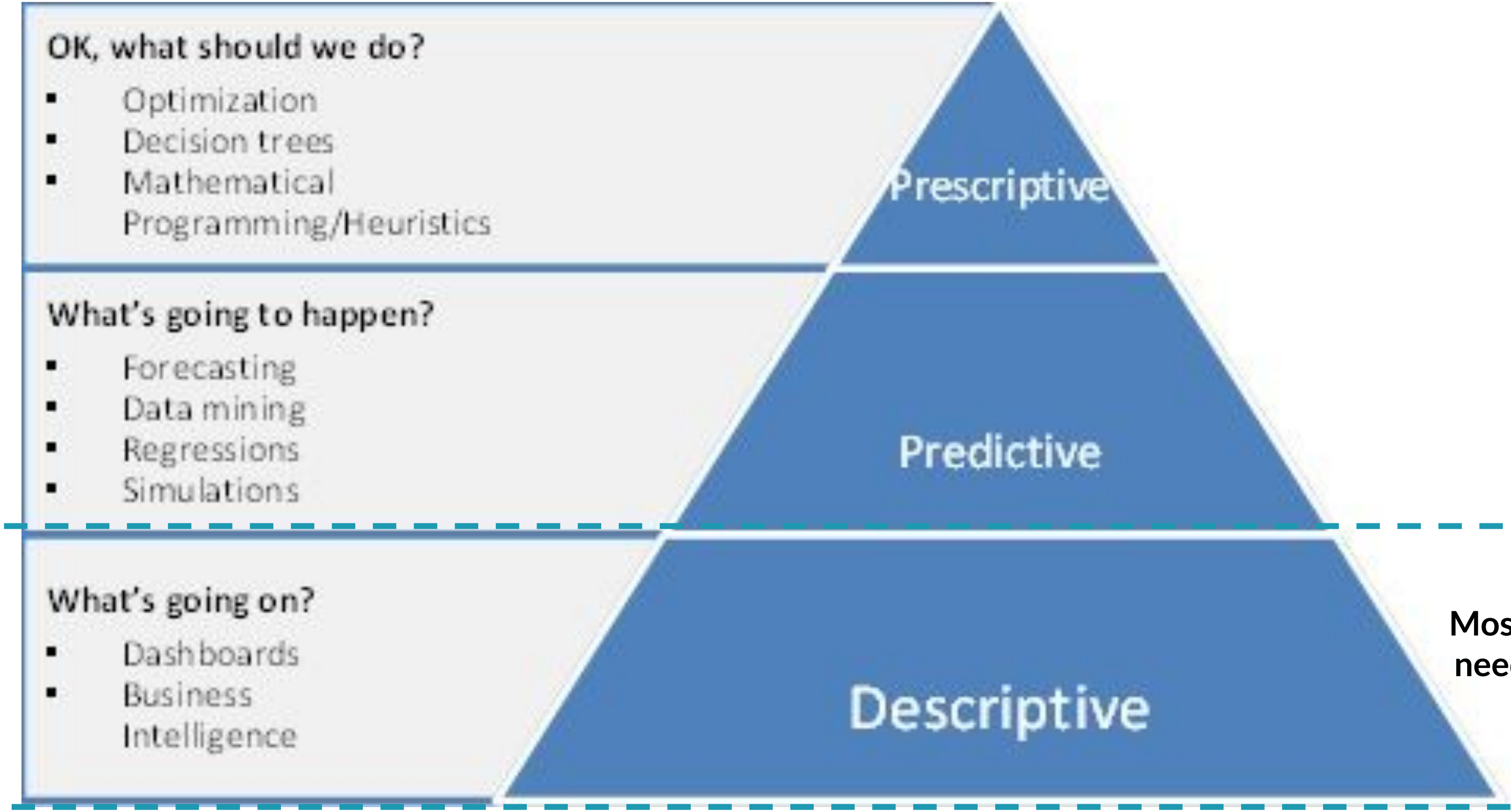


Machine learning + Exploratory data analysis

Hierarchy of Analysis

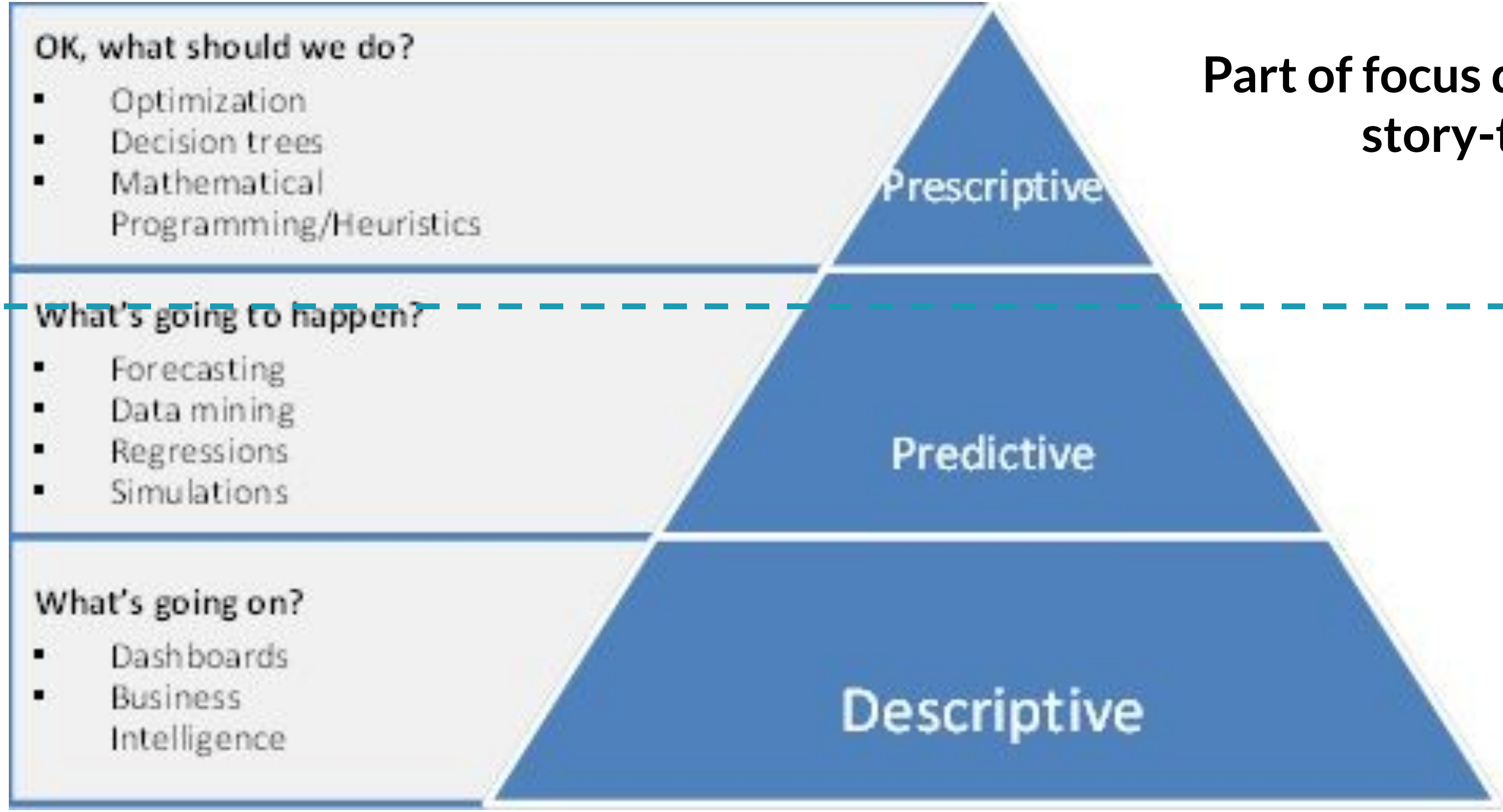


Hierarchy of Analysis



Most common need of public sector

Hierarchy of Analysis



Part of focus during story-telling

In Summary

1. Review of ANOVA
2. Isolated the impact of different races on salary data
    a. Introduction to machine learning