The Levels of Obesity in South America

Gabrielle Salamanca

1.

The Dataset

Obesity Levels

The Dataset

About

- Provided by the DSS Club
  - kaggle.com
- From:
  - Mexico
  - Peru
  - Colombia
- 2111 rows
- 17 columns
- No missing values

Contains

- Personal information
  - Age
  - Gender
  - Height
  - Weight
- Habits
  - Eating
  - Exercise
  - Smoking

The Dataset: Response Variable

Variable: Nobeyesdad

- Obesity lvl
  - Weight category
  - In kilograms

Category:

- Insufficient Weight
- Normal Weight
- Overweight Lvl I
- Overweight Lvl II
- Obesity Type I
- Obesity Type II
- Obesity Type III

The Dataset: Response Category

Insufficient Weight

- 39-65 kg
- 85.98-143.3 lbs

Normal Weight

- 42.3-87 kg
- 93.26-191.8 lbs

Overweight Lvl I

- 53-91 kg
- 116.85- 200.62 lbs

Overweight Lvl II

- 60-102 kg
- 132.28-224.87 lbs

Obesity Type I

- 75-125 kg
- 165.35-275.58 lbs

Obesity Type II

- 93-130 kg
- 205.03-286.6 lbs

Obesity Type III

- 102-173 kg
- 224.87- 381.4 lbs

The Dataset: ***NOTE!***

These categories are likely to have BMI (Body Mass Index) in mind

- Metric
  - $$\frac{weight\ (kg)}{height^2(m)} = BMI$$
- Imperial
  - $$703\left[\frac{weight\ (lbs)}{height^2(in)}\right] = BMI$$

The Dataset: Goal

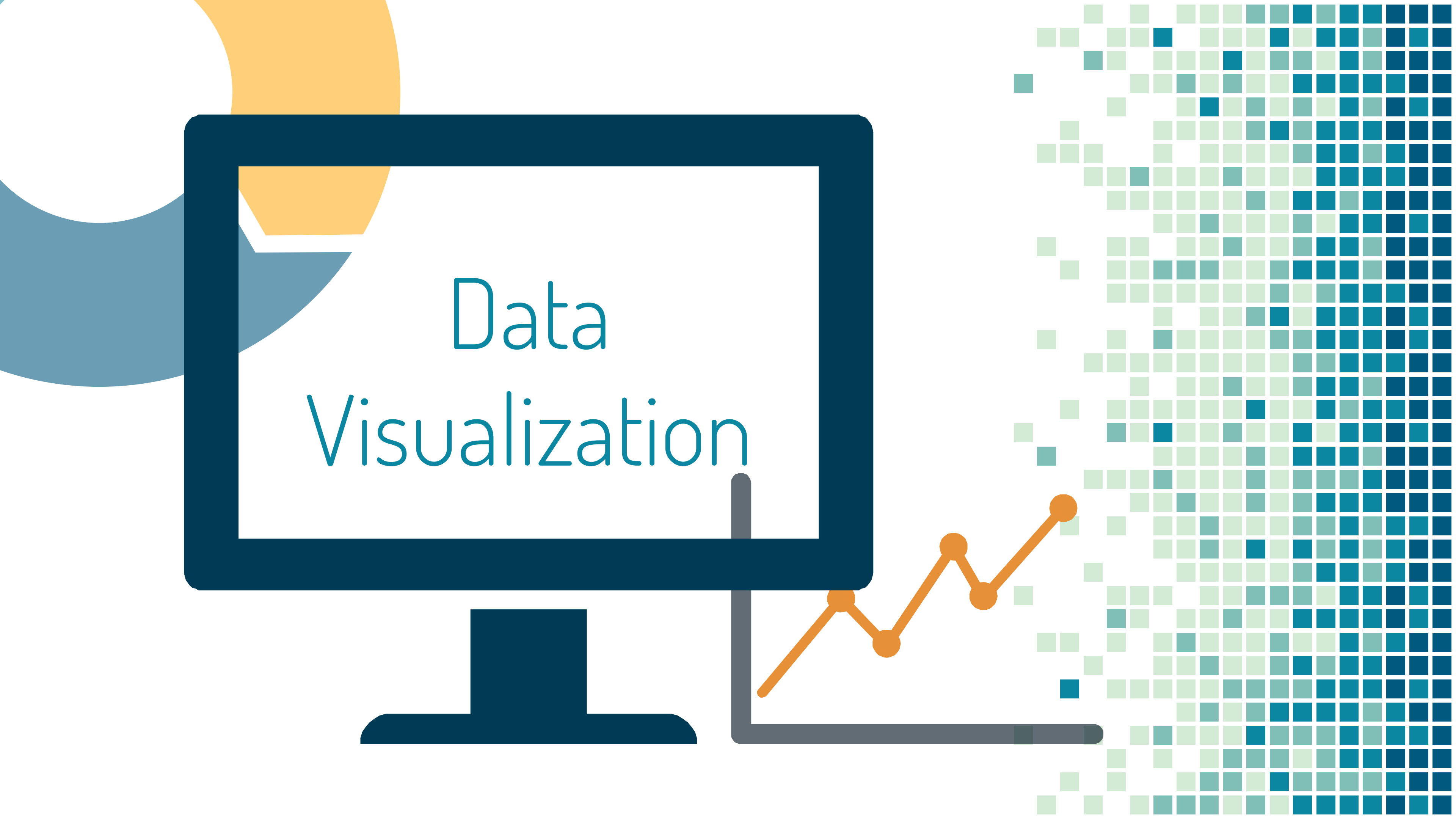
From kaggle.com:

Obesity, which causes physical & mental problems, is a global health problem with serious consequences. The prevalence of obesity is increasing steadily, & therefore, new research is needed that examines the influencing factors of obesity & how to predict the occurence of the condition according to these factors.

> “Your BMI concerns your height and weight. But are there other factors that affect your weight category?

The Dataset: Goal

- Classification
  - Find a model provides the most accurate prediction
- Response variable
  - Obesity lvl



Data Visualization

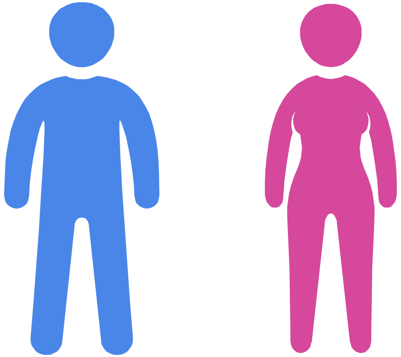
ObesityDataSet_raw_and_data_sinthetic

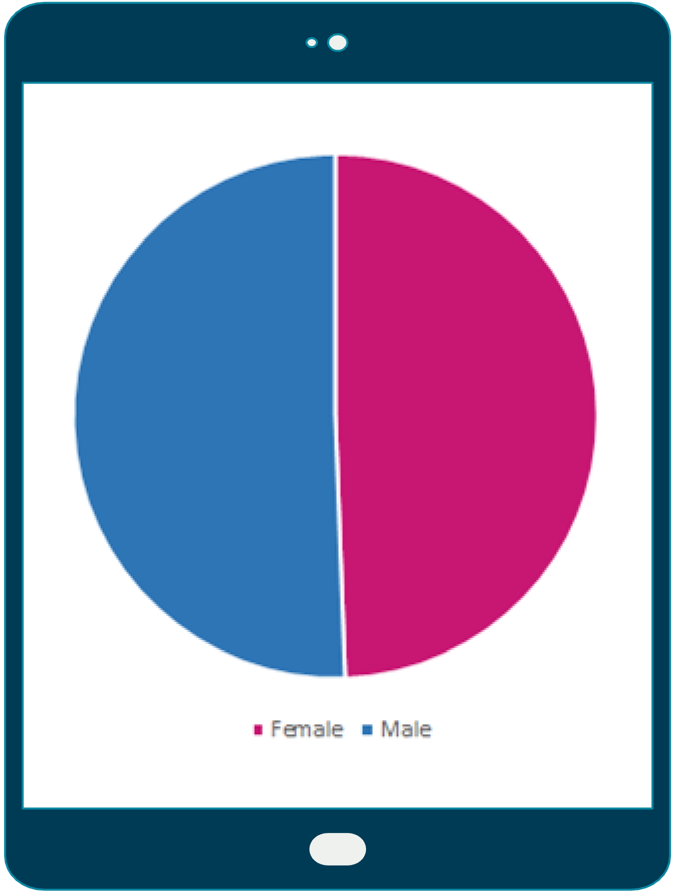
	Age	Gender	Height	Weight	Alcohol Consumption	...
1	21	Female	1.62	64.0	No	...
2	21	Female	1.52	56.0	Sometimes	...
3	23	Male	1.80	77.0	Frequently	...
4	27	Male	1.80	87.0	Frequently	...
5	22	Male	1.78	89.9	Sometimes	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

The Gender Pie

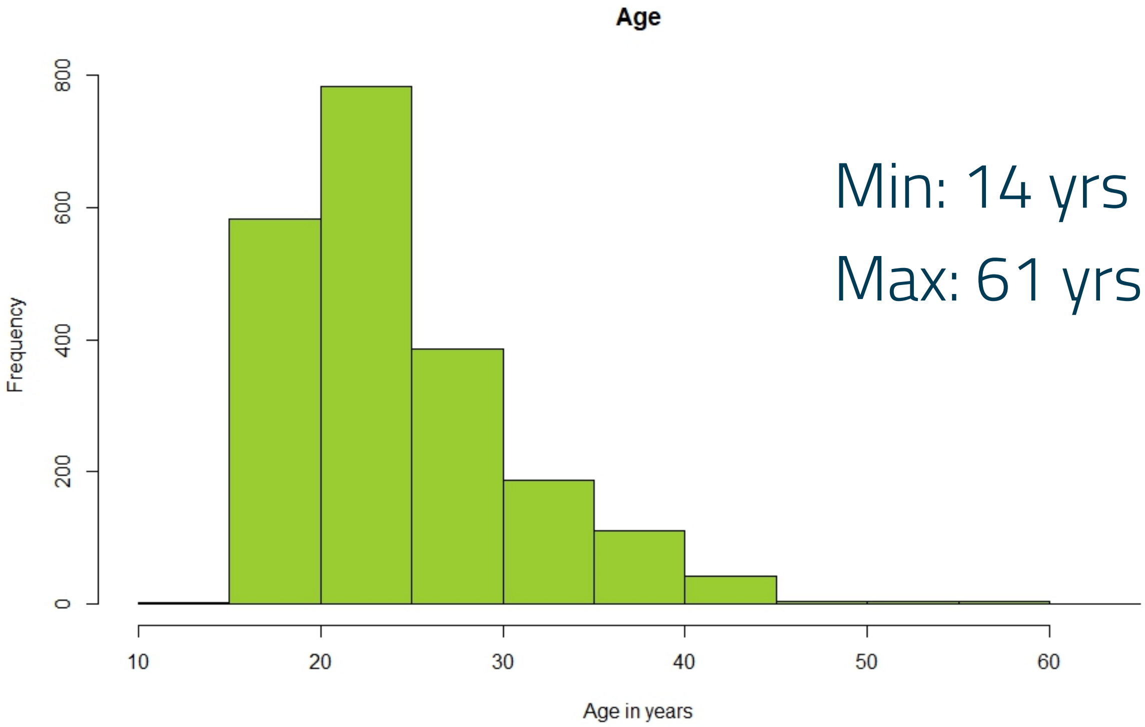
2111 participants

- 1043 females
- 1068 males





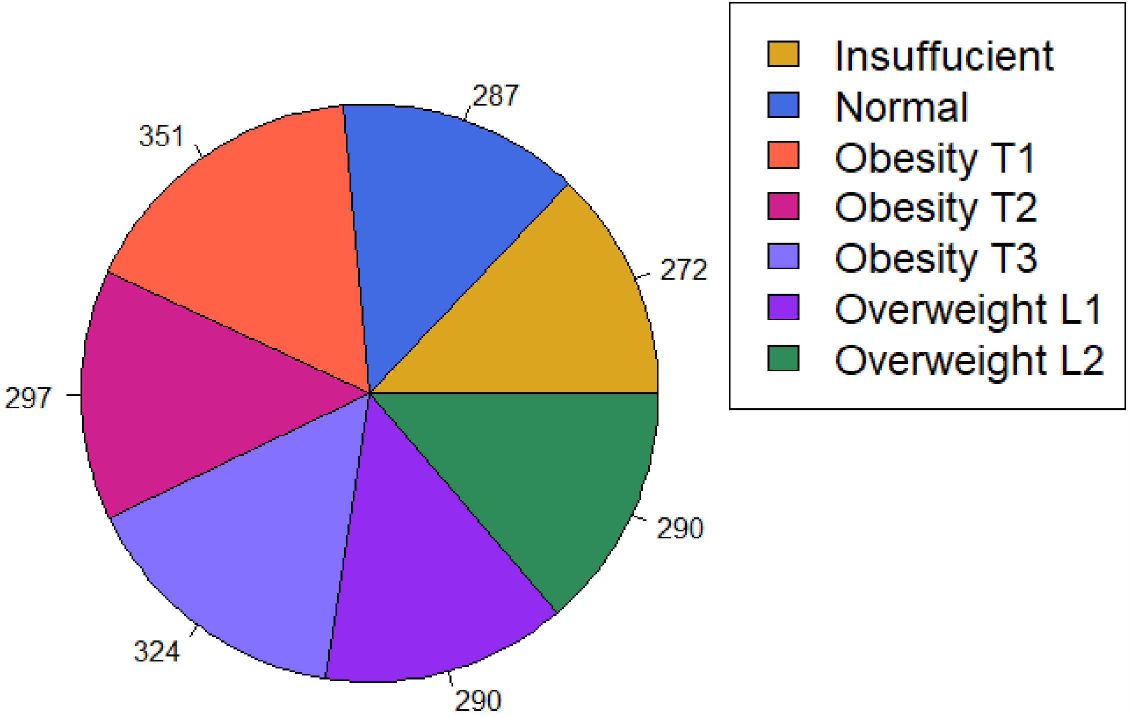
Histogram: Age



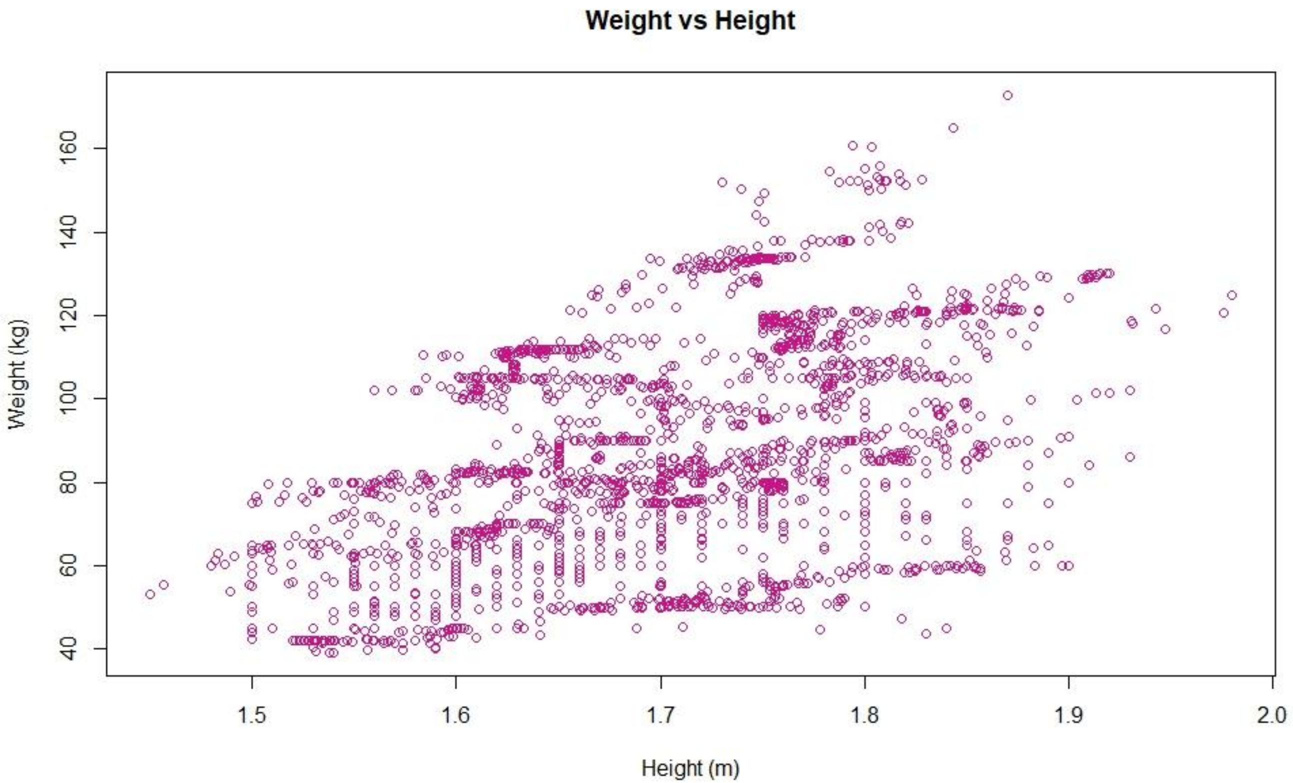
Min: 14 yrs old

Max: 61 yrs old

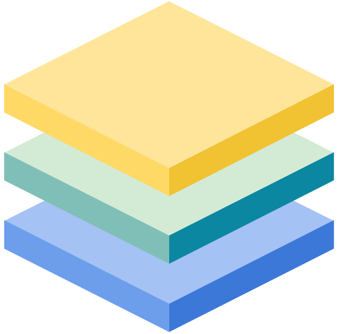
Weight Categories



Weight vs Height



2.
Cleanup, Organization, & Such



Additional steps before continuing?

Data Cleanup

None!

- No missing values
- No notable outliers

Consideration:

- Removing category
  - Insufficient Weight

If looking at obesity lvls

- Concern only for the obese/overweight
- Normal for baseline

Data Consideration

New dataset

- Removed "insufficient weight" category
- Separate train & test data

Results:

- **Error:**
  - All arguments must have the same length
- **Warning:**
  - longer object length is not multiple of shorter object length

Organization

Original columns	Renamed columns
CALC	Alcohol Consumption
FAVC	High Caloric Food Consumption
FCVC	Vegetable Consumption
TUE	Screen Time
CAEC	Snacking
Nobeyesdad	Obesity lvl

3. Fits & Methods

What has been done?

Model Fits: GLM

Data Splitting

- 80% training
- 20% test

1. Manually go through fits for all 16 vars
   a. Linear → Max
2. Max variable fits

summary()

- Find most significant vars
- Use a variation of those to find best fit

Model Fits: GLM

anova()

- Didn’t give the significant vars
- Warnings
  - Fitted probabilities numerically 0 or 1 occurred
  - Algorithm didn’t converge

chisq.test()

- Error
  - ‘List’ obj can’t be coerced to type ‘double’

Model Method: Best Subset Selection

Best Subset Selection

- Adjusted $R^2 = 16$ vars
- Mallow's $C_p = 16$ vars
- $BIC = 11$ vars

Opinion

- summary(16 variables)
  - No significant p-values
    - 0.982 - 1
- BIC
  - More reasonable

Model Fits: 1st Version

glm(Obesity.Lvl~)

- Age, Gender
- Weight, High.Caloric.Food.Consumption
- Main.Meal.Consumption, Calorie.Count
- Water.Consumption
- Family.History.Overweight
- Exercise.Activity, Screen.Time
- Snacking

summary(): $\alpha > 0.05$

- Gender
- High.Caloric.Food.Consumption
- Calorie.Count
- Family.History.Overweight
- Exercise.Activity
- Screen.Time

Model Fits: 2nd Version

glm(Obesity.Lvl~)

- Age
- Weight
- Main.Meal.Consumption
- Water.Consumption
- Snacking

summary() Results, $\boldsymbol{\alpha = 0.05}$

- All p-values significant
  - But Snackingno = 0.155470
- AIC = 330.36
- $\chi^2_{1687-1680}$ = null dev - residual dev
  - 1306.57 - 314.36 = 992.21
  - p-value = .00001

Model Fit & Methods

Chosen fit

- glm.fit2 = glm(Obesity.Lvl ~ Age + Weight + Main.Meal.Consumption + Water.Consumption + Snacking)

Methods

- QDA
- LDA

Method: Linear Discriminant Analysis

Classification

LDA

Supervised learning algorithm specifically designed for classification tasks, aiming to identify linear combo of features that optimally segregates classes w/in dataset

Source:

https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/

Method: Quadratic Discriminant Analysis

Classification

QDA

Similar to LDA

- Relaxed assumption
  - Mean & coV of all classes are equal
- Calculation done separately ∀ class

Source:

https://www.geeksforgeeks.org/quadratic-discriminant-analysis/

Model Fit: Prediction

Glm.fit2 = 0%

	Insufficient	Normal	Ob T1	Ob T2	Ob T3	Ow L1	Ow L2
No	**44**	**11**	**0**	**0**	**0**	**0**	**0**
Yes	**9**	**39**	**71**	**57**	**83**	**58**	**51**

Model Fit: Method Testing

QDA

- Error
  - Rank deficiency in group Obesity Type 3

LDA

- No errors
- 59.81087%

Model Fits: 3rd Version

1. Run new fit I've used previously
2. Use summary() to see best variables
   1. Height
   2. Weight

glm.fit3 = glm(Obesity.Lvl ~ Weight + Height)

- Runs w/ no errors when running the other methods

Model Methods: QDA

	Insufficient	Normal	Ob T1	Ob T2	Ob T3	Ow L1	Ow L2
Insufficient	57	3	0	0	0	0	0
Normal	0	45	0	0	0	0	0
Ob T1	0	0	73	0	0	0	1
Ob T2	0	0	0	55	6	0	0
Ob T3	0	0	0	1	63	0	0
Ow L1	0	2	0	0	0	46	6
Ow L2	0	0	2	0	0	2	61

Model Methods: LDA

	Insufficient	Normal	Ob T1	Ob T2	Ob T3	Ow L1	Ow L2
Insufficient	48	6	0	0	0	0	0
Normal	9	33	0	0	0	0	0
Ob T1	0	0	71	0	0	0	1
Ob T2	0	0	2	56	13	0	0
Ob T3	0	0	0	1	56	0	0
Ow L1	0	11	0	0	0	46	6
Ow L2	0	0	2	0	0	2	61

Model Methods: Results

Prediction

- 0%

QDA

- 94.56265%

LDA

- 87.70686%

4. Improvements

That said...

If I had more time…

1. Further exploration on best fit
   a. Esp making it consistent
   b. summary(fit) result changes every time I go back to R
2. Figure out the prediction table with the fit
3. Apply more methods

CREDITS

Presentation template

- SlidesCarnival

Dataset

- Data Science Society
- kaggle.com

THANKS!

Any questions?
Any suggestions?