Prevalence of obesity based on dietary/work practices, genetics and physical condition

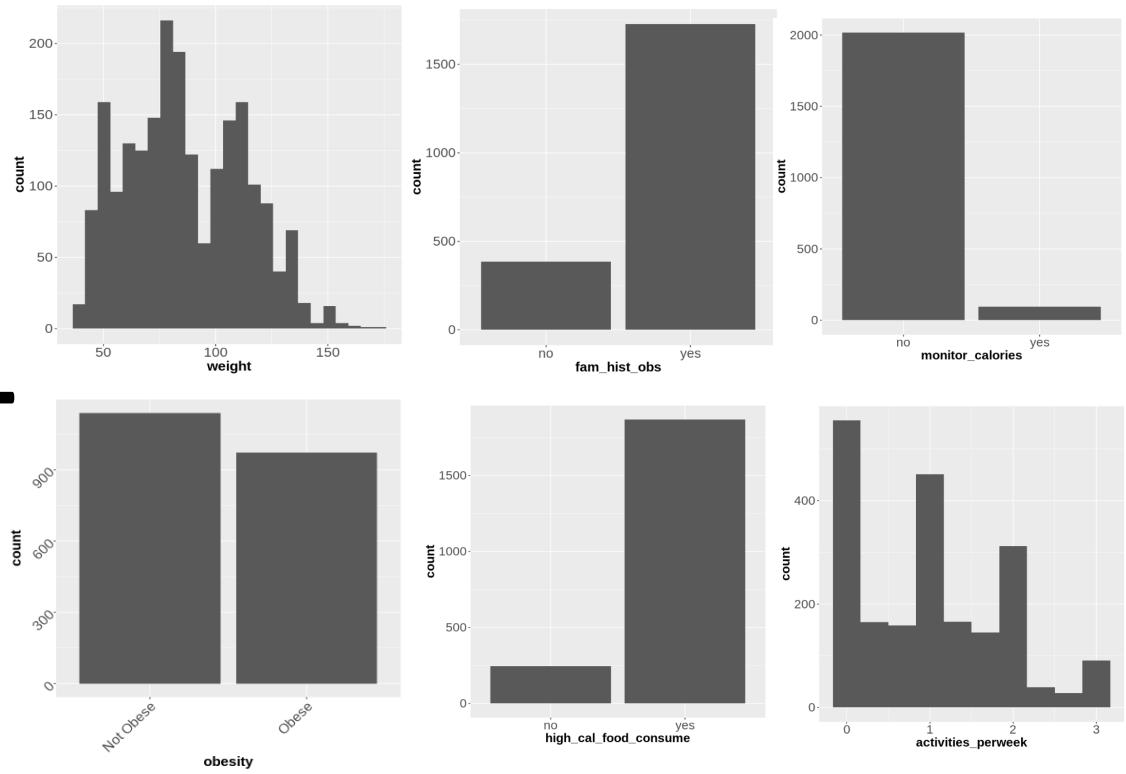
Source: https://archive-beta.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

STEP-1 Data loading, cleaning & display: 2111 rows/17 columns; renamed columns, converted characters to factors and reduced obesity levels from 7 to 2, “Not obese” or “Obese”

	gender	age	height	weight	fam_hist_obs	high_cal_food_consume	meals_with_vegetables	meals_a_day	food_between_meals	smoke	water_consumption	monitor_calories	activities_perweek	time_tech_uses	alcohol_intake	transportation_usage	obesity
	<fct>	<dbl>	<dbl>	<dbl>	<fct>	<fct>	<dbl>	<dbl>	<fct>	<fct>	<dbl>	<fct>	<dbl>	<dbl>	<fct>	<fct>	<fct>
1	Female	21	1.62	64.0	yes	no	2	3	Sometimes	no	2	no	0	1	no	Public_Transportation	Not Obese
2	Female	21	1.52	56.0	yes	no	3	3	Sometimes	yes	3	yes	3	0	Sometimes	Public_Transportation	Not Obese
3	Male	23	1.80	77.0	yes	no	2	3	Sometimes	no	2	no	2	1	Frequently	Public_Transportation	Not Obese
4	Male	27	1.80	87.0	no	no	3	3	Sometimes	no	2	no	2	0	Frequently	Walking	Not Obese
5	Male	22	1.78	89.8	no	no	2	1	Sometimes	no	2	no	0	0	Sometimes	Public_Transportation	Not Obese
6	Male	29	1.62	53.0	no	yes	2	3	Sometimes	no	2	no	0	0	Sometimes	Automobile	Not Obese

STEP-2 Data exploration: Count histogram

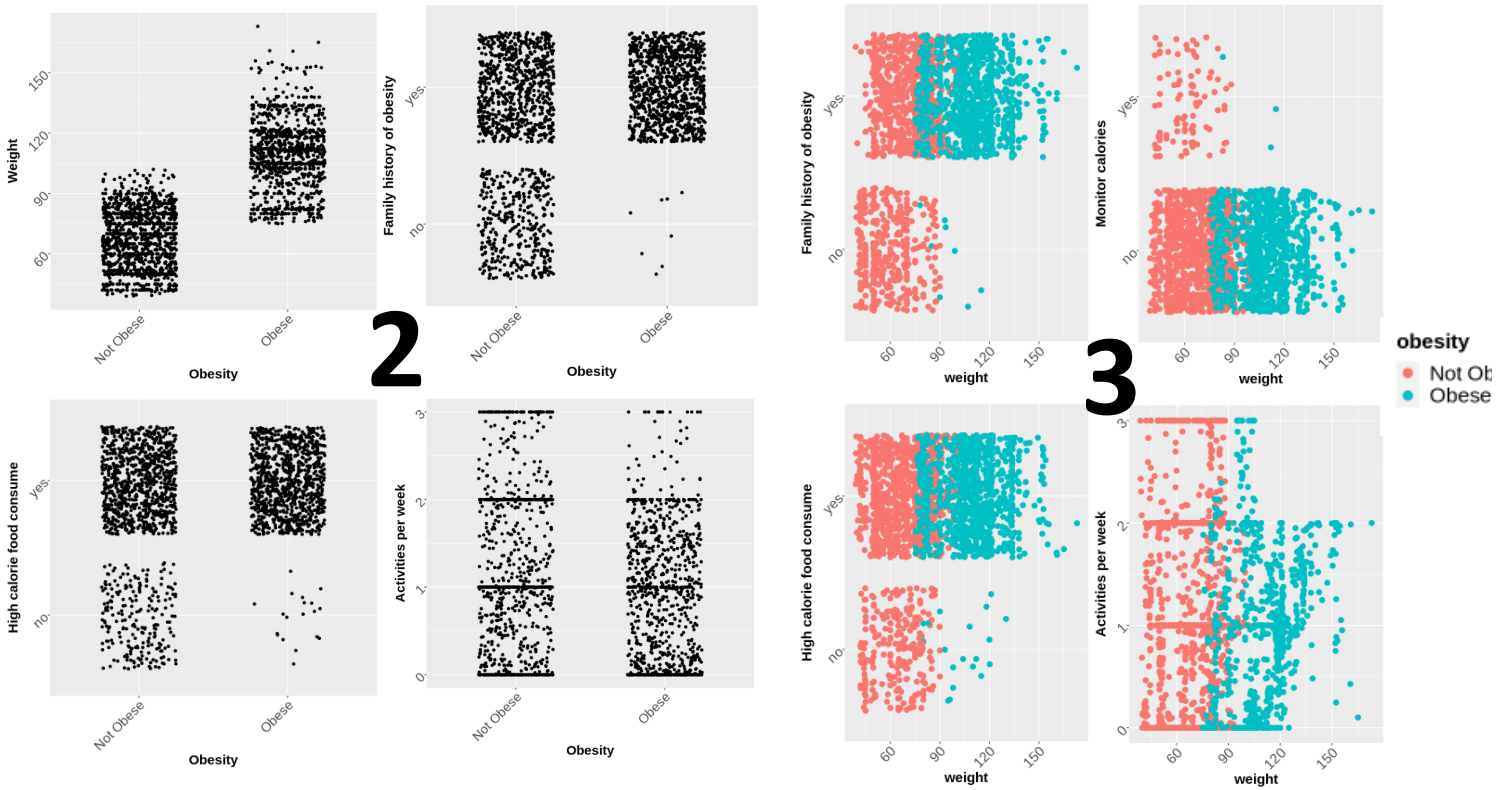
1

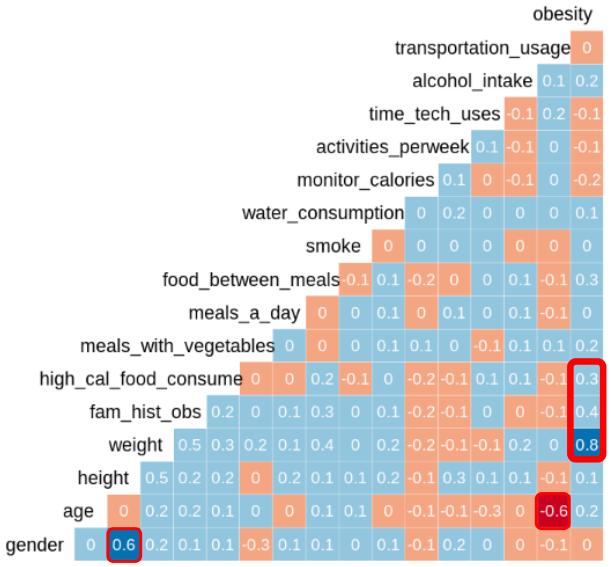


Few observations from Panel 1:

- Weight shows a bimodal distribution.
- Number of non-obese to obese people is not much different.
- On the contrary, the number of people who have family history of obesity, who consume high calorie foods, and who monitor calories is vastly different from those who do not.
- Activities per week is a discrete numerical variable.

STEP-2 (Con.) Jitter plots for bivariate relation and trivariate relations, correlation plot

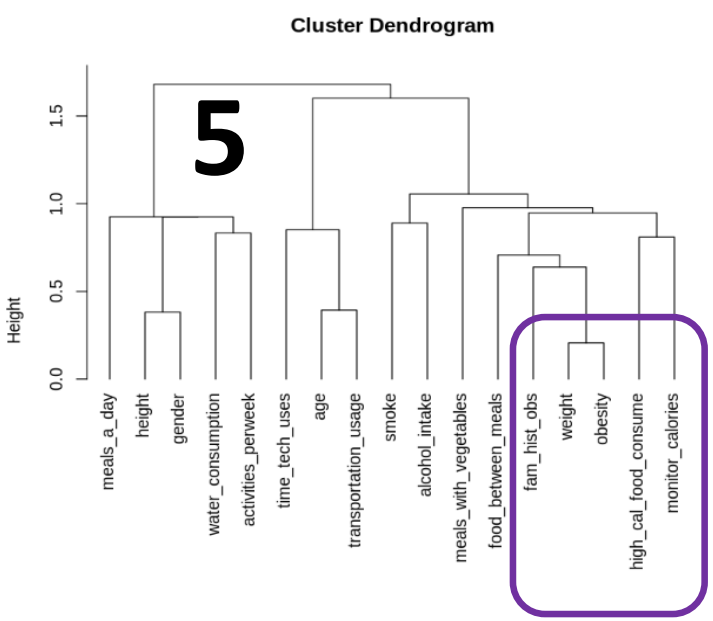


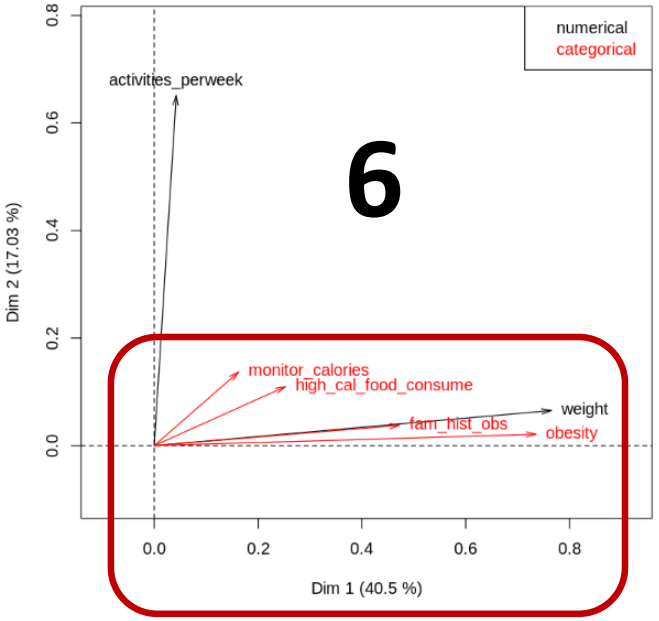


Few observations

- **Panel 2 (left black plot) :** Among bivariate jitter plots (data distribution), weight of obese people is clearly different from non-obese; people with no family history of obesity are mostly non-obese; people who do not eat high calorie food are mostly non-obese; obese people have usually lower values of activities per week. Obese people have more weight.
- **Panel 3 (middle color plot):** Among trivariate jitter plots (data distribution), obese people (blue) have more weight. Also, obese people have higher history of family obesity cases, do not monitor calories while eating, consume high calorie foods and tend to have lower activities per week. Panel-3 and Panel-2 tend to provide us with similar conclusions but Panel-3 does better job in visualizing the differences among obese and non-obese people.
- **Panel 4 (right):** A correlation of variables suggested positive correlation between obesity and weight (0.8), family history of obesity (0.4) and high calorie food consume (0.3).

STEP-3 Variable clustering and PCA for mixed data for variable relationship





Few observations from Panel 5 & 6: Both clustering of variables and PCA with reduced number of variables suggest that obesity, family history of obesity, high calorie food consume, minor calories and weight influence dimension1 of PCA which explains 41% of variation in data.

STEP-4 Logistic regression in predicting obesity classification & evaluate model fit

- Split data into 66% training & 34% testing dataset; use training data to generate optimized model with weight, family history of obesity and activities per week as predictor variables

```
split = sample.split(df$obesity, SplitRatio = 0.66)
train = subset(df, split == TRUE)
test  = subset(df, split == FALSE)
```

```
model3 <- glm(obesity ~ weight + fam_hist_obs + activities_perweek,
family=binomial(link='logit'), maxit=100,data=train[,c(4,5,13,17)])
```

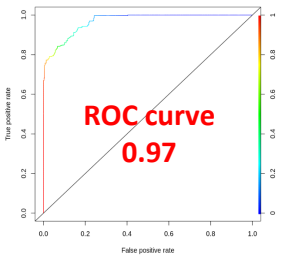
```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -17.26230    1.19844 -14.404  < 2e-16 ***
weight             0.17983    0.01159  15.513  < 2e-16 ***
fam_hist_obsyes    2.02608    0.73981   2.739 0.006169 **
activities_perweek -0.45219    0.12385  -3.651 0.000261 ***

    Null deviance: 1923.81  on 1393  degrees of freedom
Residual deviance:  572.21  on 1390  degrees of freedom
AIC: 580.21
```

- Evaluated model fit for predicting obesity classification in test dataset: Obtained high accuracy, sensitivity, specificity, precision and ROC metric (0.84-0.97)

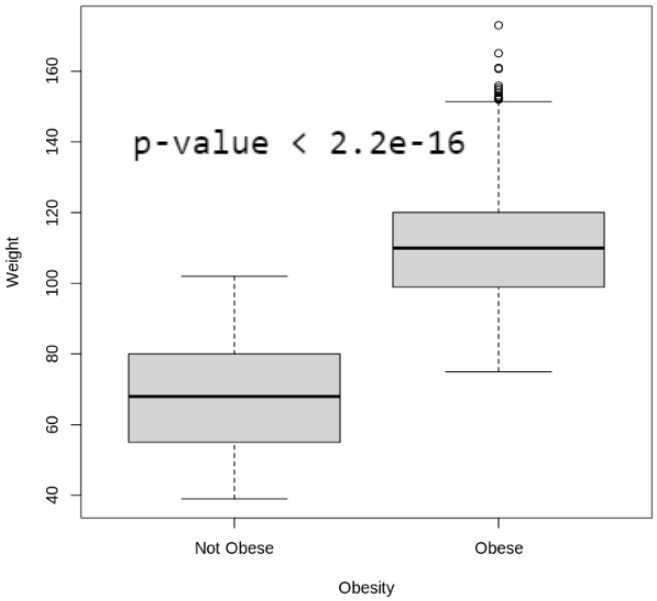
```
probs1 = predict(model3, type = "response", newdata=test)
preds1 = ifelse(probs1 > 0.5,1,0)
```

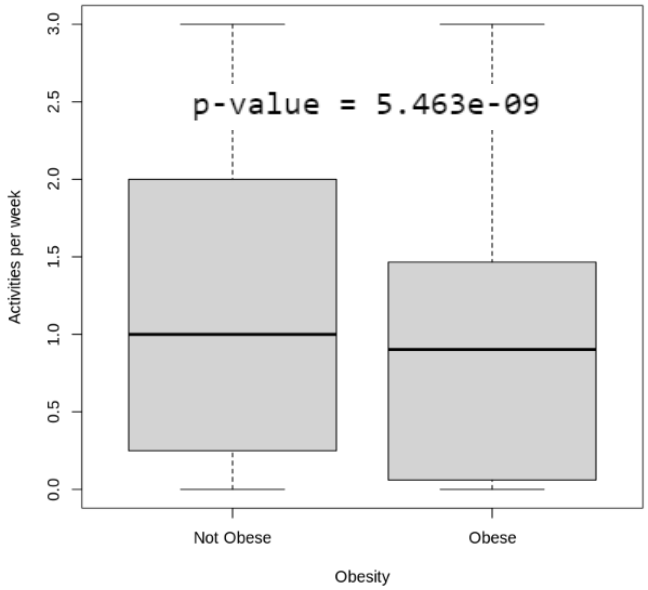
```
table(test$obesity,preds1>0.5)
```

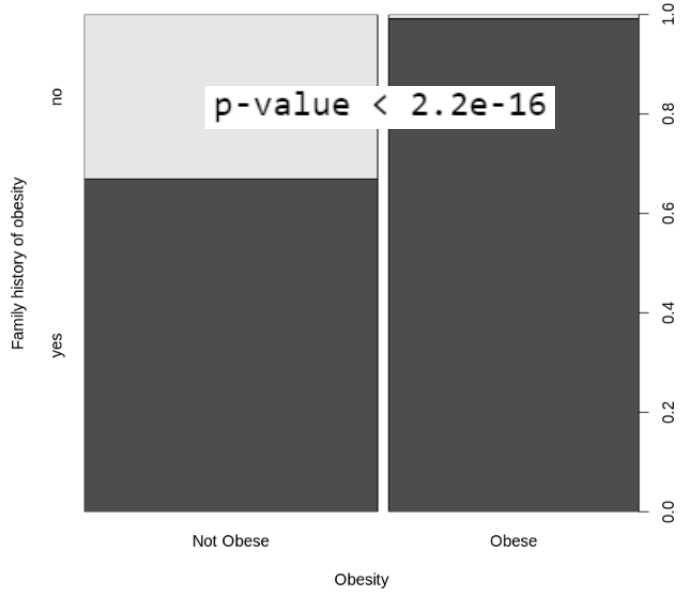
```
           FALSE TRUE
  Not Obese   361   26
  Obese        52  278

[1] "Accuracy = 0.891213389121339"
[1] "Sensitivity = 0.842424242424242"
[1] "Specificity = 0.9328165374677"
[1] "Precision = 0.9144736842105263"
```



STEP-5 Statistical tests based on obesity classification in weight, activities per week and family history of obesity







Weight vs Obesity

- Unequal variance
- Non-normality
- Wilcoxon test
- Null hypothesis rejected
- Mean weights of obese and non-obese people are statistically significantly different.

Activities per week vs Obesity

- Unequal variance
- Non-normality
- Wilcoxon test
- Null hypothesis rejected
- Mean activities per week of non-obese people is significantly greater than obese people.

Family history of obesity & Obesity

- Both categorical variables
- Chi-square test
- -Fisher Exact test
- Null hypothesis rejected
- Ratio of people with a family history of obesity to those who do not is significantly higher in obese than non-obese people.

Summarizing, a person has more chance of being classified as obese if the weight is very high, has lower activity per week or has high family history of obesity. Thus, physical condition, work practices and genetics play a role in predicting obesity.