Data Quest – Unlocking the Power of Data

(Case Study Challenge '23)

Track	***From fork to fitness**: Relationship between Eating Patterns and Obesity*
College/University	*Savitribai Phule Pune University*
Software Used	*R*

Student Name	Email ID	PG Course	Current semester
Ayshik Neogi	ayshikneogi13@gmail.com	MSc Statistics	4th
Darshan Mali	malidarshan3@gmail.com	MSc Statistics	4th
Saurav Jadhav	sauravjadhav698@gmail.com	MSc Statistics	4th

Instructions:

(a) The below slides are mandatory. Addition of slides are allowed. The final PPT should not exceed more than 8 slides.

(b) Use bullet pointers/graphs to make your presentation more concise and effective.

What is your data speaking?

1. Majority of the teens under study are Asian.

2. The range of weight for the teens is 63 to 135 indicating teens are somewhat on the higher side of the weight.

3. Different race of a subject has been recorded from different countries(Mutually Exclusive).

4. The empirical distribution of BMI and weight of the teens are multimodal which indicates that it is more likely we have groups and subgroups within the data based on different characteristics.

5. Performing PCA to analyze the TestQ, revealing the presence of three distinct clusters. One cluster and a group of two clusters are formed and separated based on treatment. Moreover, two clusters with the same treatment are differentiated by income criteria.

Analysis & Visualization-1

Information Value (IV)

X (Categorical)	Y(Binary)	IV	Results
ECS	Wt_enc	2.813419	Highly Predictive
UCS	Wt_enc	2.575849	Highly Predictive
Visit	Wt_enc	1.56475	Highly Predictive
ECS	Treatment	0.9500519	Highly Predictive
UCS	Treatment	0.4135771	Highly Predictive
RCS	Wt_enc	0.3414344	Highly Predictive
Treatment	Wt_enc	0.3128151	Highly Predictive
Race	Wt_enc	0.008093469	Not Predictive
Age	Wt_enc	0.007317082	Not Predictive
RCS	Treatment	0.000576028	Not Predictive
Gender	Wt_enc	0.000576028	Not Predictive

Table: 1

- IV is a good measure to quantify predictive power of a independent feature on a target variable(Binary Variable).
- Here we have transformed following continuous variables into binary variables with median as cutoff for evaluating IV. UCS, ECS, RCS, Age, Weight in Table 1.
- Here, we can observed in table 1 that ECS, UCS has higher IV with target weight and Treatment. They may accurately predict Treatment.

Analysis & Visualization-2

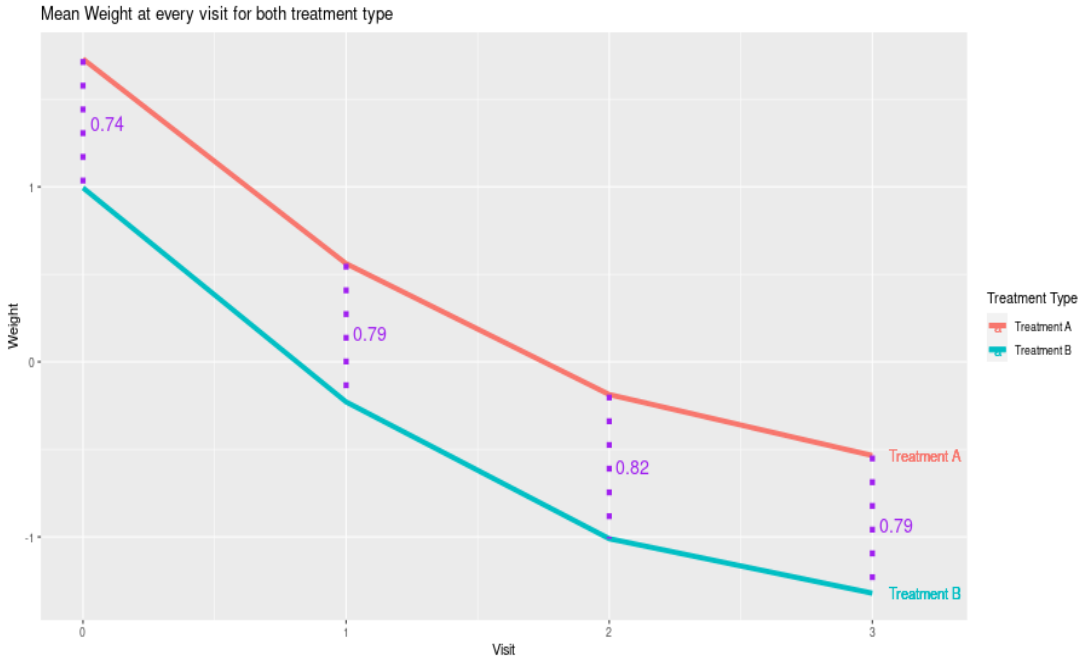

Fig 1

In Figure 1, observing the funnel-like shape is an indication of a larger fall in weight for those teens who were given treatment B. Hence we consider treatment B as our Testa, the magical pill.

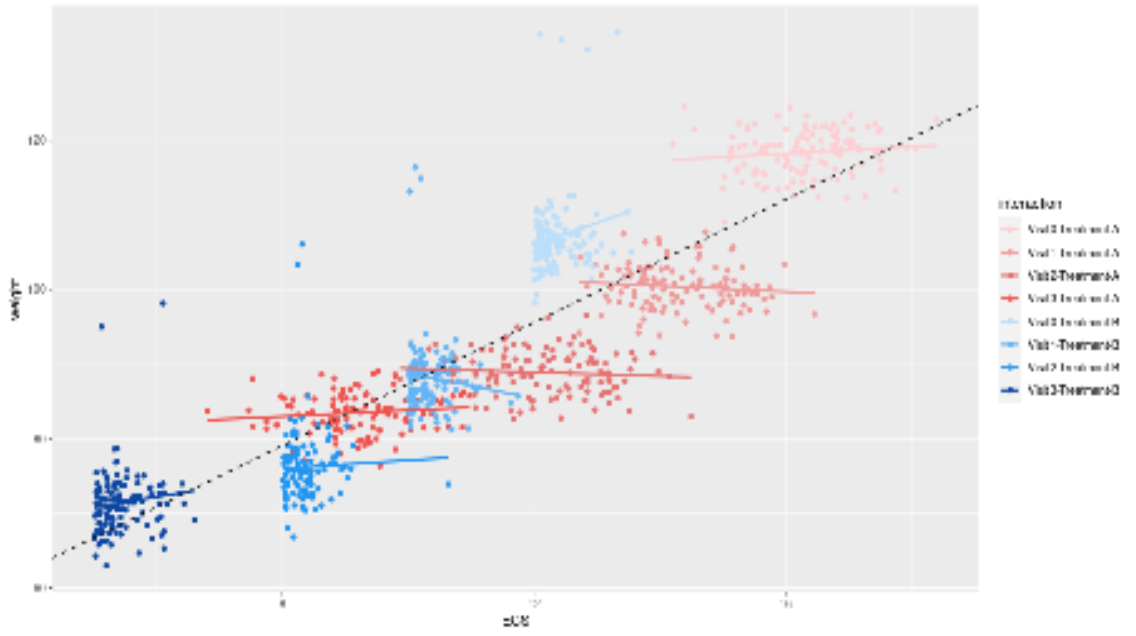

Fig 2

In figure 2, from the dotted line one can interpret ECS shares a linear relationship with Weight and with a positive slope. When we consider Visits and types of treatment, we see a different picture. In most cases, the slopes are not significant, ECS and Weight are likely to not share a relationship.

Analysis & Visualization-3

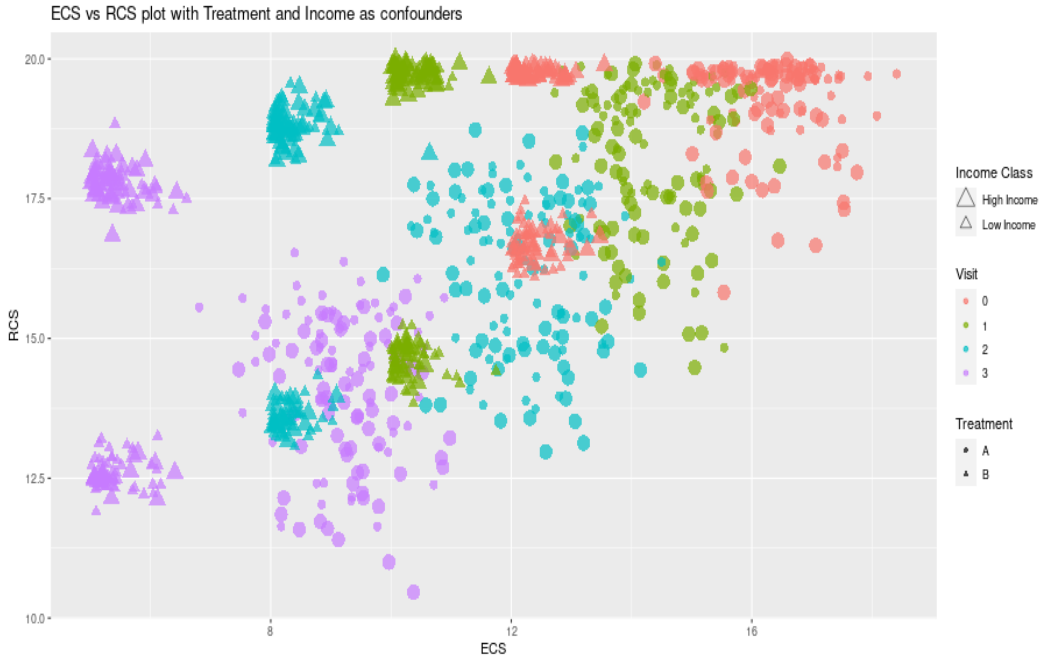

Fig 3

	RCS	ECS	UCS	Weight
RCS	2.398	0.738	0.379	0.610
ECS	0.738	2.341	0.739	1.145
UCS	0.379	0.739	1.505	0.794
Weight	0.610	1.145	0.794	2.233

Table 2

If we observe the scatter plot in fig 3, we see some pattern for the distribution of the ECS and RCS values. Those teens who were given treatment B and have a higher RCS value also tend to have higher median family income similarly, those with a lower median family income tend to have a lower score of RCS value.

From the above mutual information table we can see that:

1. ECS share a strong relationship with Weight.
2. ECS share a strong relationship with UCS.
3. ECS shares a strong relationship with RCS.

Method used and Justification-1

- 1.To decide which treatment is our Testa, we use a unique approach. At every visit, we consider the mean weight for all those patients who were given treatment A. We do a similar thing for those who were given treatment B.

  2. Now we calculate the distance between both the treatment means of the weight for every visit. We compare the distance of the last visit to that of the first visit.

  3. Why we haven't considered BMI?
  When we calculated height using BMI and weight, we realized that mean height was falling after every visit, which seems to be an anomaly since the people under the study are teenagers, the height should have been at least constant.

- We use mutual information (MI) to identify which of the features shares a strong relationship with one another. The reason to use MI is it helps to quantify the strength of linear as well as non-linear relationship.

- Note: First visit gender entry for an individual is consider for further analysis for that teen where there's an anomaly.

Method used and Justification-2

- After observing that weight and ECS are relatively strongly related, we plot the scatter diagram between them, We suspect that the relationship between them is influenced by other features and we identify those as treatment type and Visits. These variables are known as confounders. We fit separate regression lines to the observations based on visits and treatment type.

- We observe ECS and RCS do share a relatively strong relationship. And similar to the above case here too we do the scatter plot. In this case, we found out not only treatment and visits have the influence but income is also an influential variable.

- IV measures the amount of uncertainty or randomness in a variable. It is commonly calculated based on the distribution of the target variable (e.g., whether an event occurred or not) across different levels or bins of the predictor variable.
  In our case taking median as a cutoff for transforming continuous variables into binary variables gives us logical and sensible feature selection for further model building.

Interpretation & Conclusion

1. From figure 2 with treatment and visits we can see that there's a reduction in weight and the tendency to overeat in response to negative emotions.

2. From figure 3, we see that those with a higher median family income tend to have a higher score of RCS value. The probable reason for this might be due to the fact that those with a higher income are more likely to have unrestricted access to food.

3. And from visit to visit we also see a drop in the ECS as well as in the RCS values, which can be a result of treatment triggering the mental response.

4. As shown in Table 1: ECS, RCS has higher IV hence have higher predictive power. These scores can be used to increase accuracy and precision in logistic regression model with transformed weight and treatment as target variable individually.

Recommendations/Scope

1. Based on the analysis, we can see that visits and treatment has an effect on the TestQ score, which indicates that there's a trigger in the mental responses and we can extend this idea incorporating it in the daily life and address the obesity from the roots.

2. In figure 2 we observed the Simpson's paradox. We can incorporate the idea of causal inference and gather deeper insights about the eating behavior and treatment effect.

3. From the BMI we can find the teens who are overweight, or obese or healthy at every visit. We can estimate the Transition Probability Matrix (TPM) and see from visit to visit how many individuals are transitioned from overweight or obese to healthy category for both the treatments.

4. As features with higher IV can also be in features selection, we can build model around those features. Especially linear model or logistic regression will give us good results.

5. Evaluating healthiness of an individual basis on BMI, weight and other given feature is not enough. More health information of an individual can give us good results. e.g. Body response after running for 5 min, heart is one of the most affected body organ in obese people so data related to heart should be consider to evaluate health status of an individual.