Data Science Bootcamp

Predicting Primary School Achievement with School-Related and Socioeconomic Factors

Abou DIENG & Sidi DOUMBOUYA

5 juin 2020

Context

- The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ).
  - Rethinking education policies
  - Guiding reforms
  - Developing educational programmes
  - Managing challenges to the quality of education

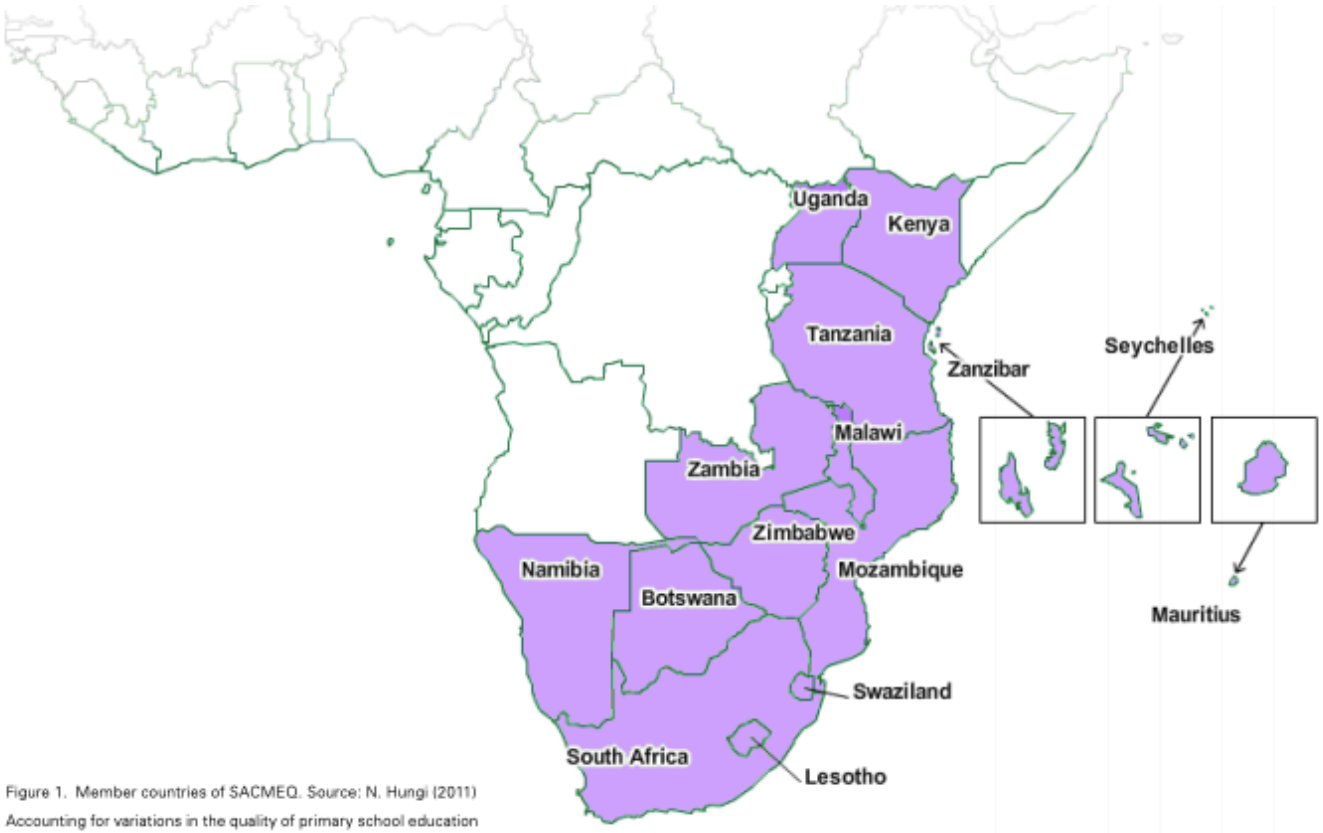
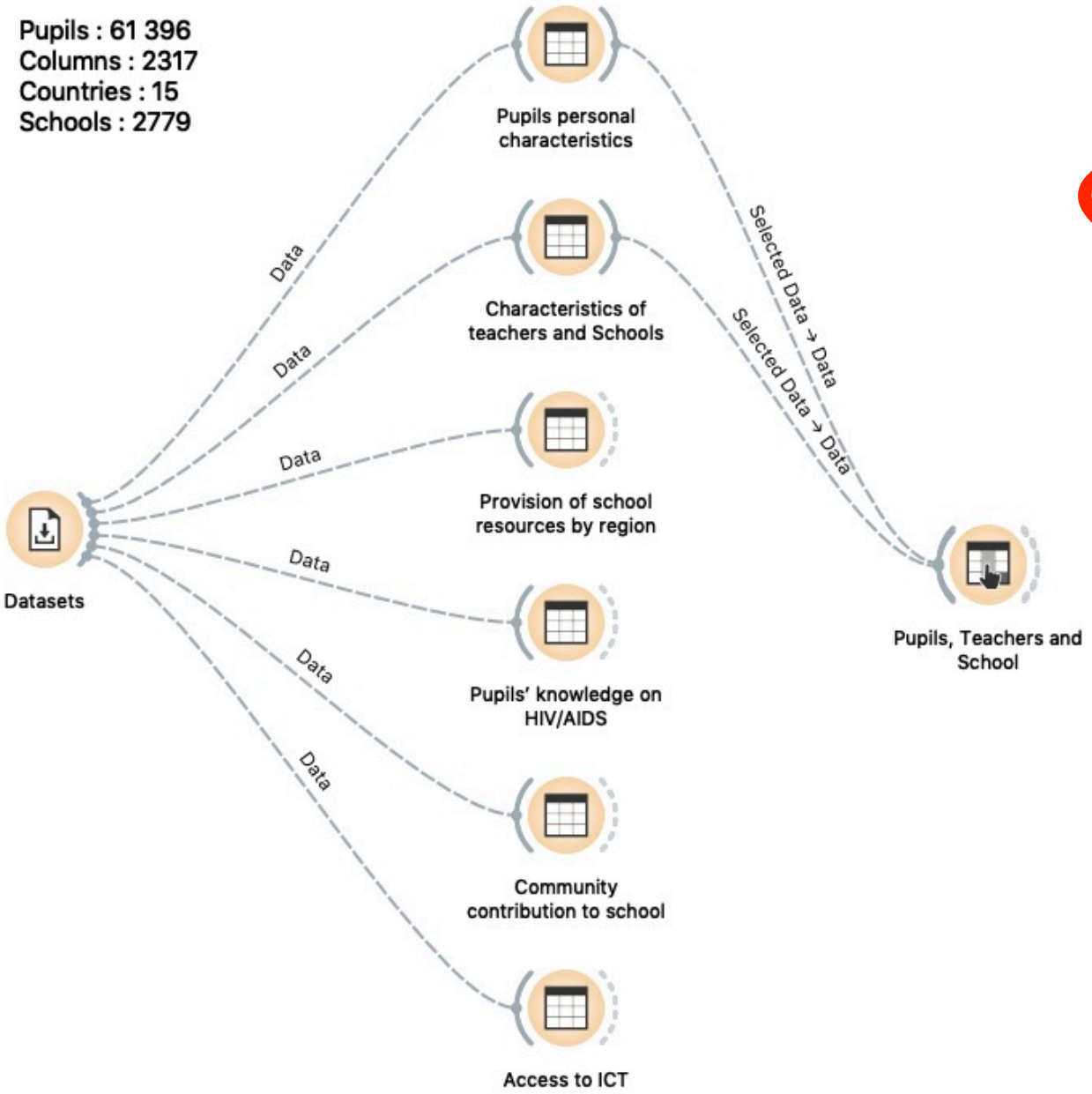


Figure 1. Member countries of SACMEQ. Source: N. Hungi (2011)
Accounting for variations in the quality of primary school education

Contents

Dataset

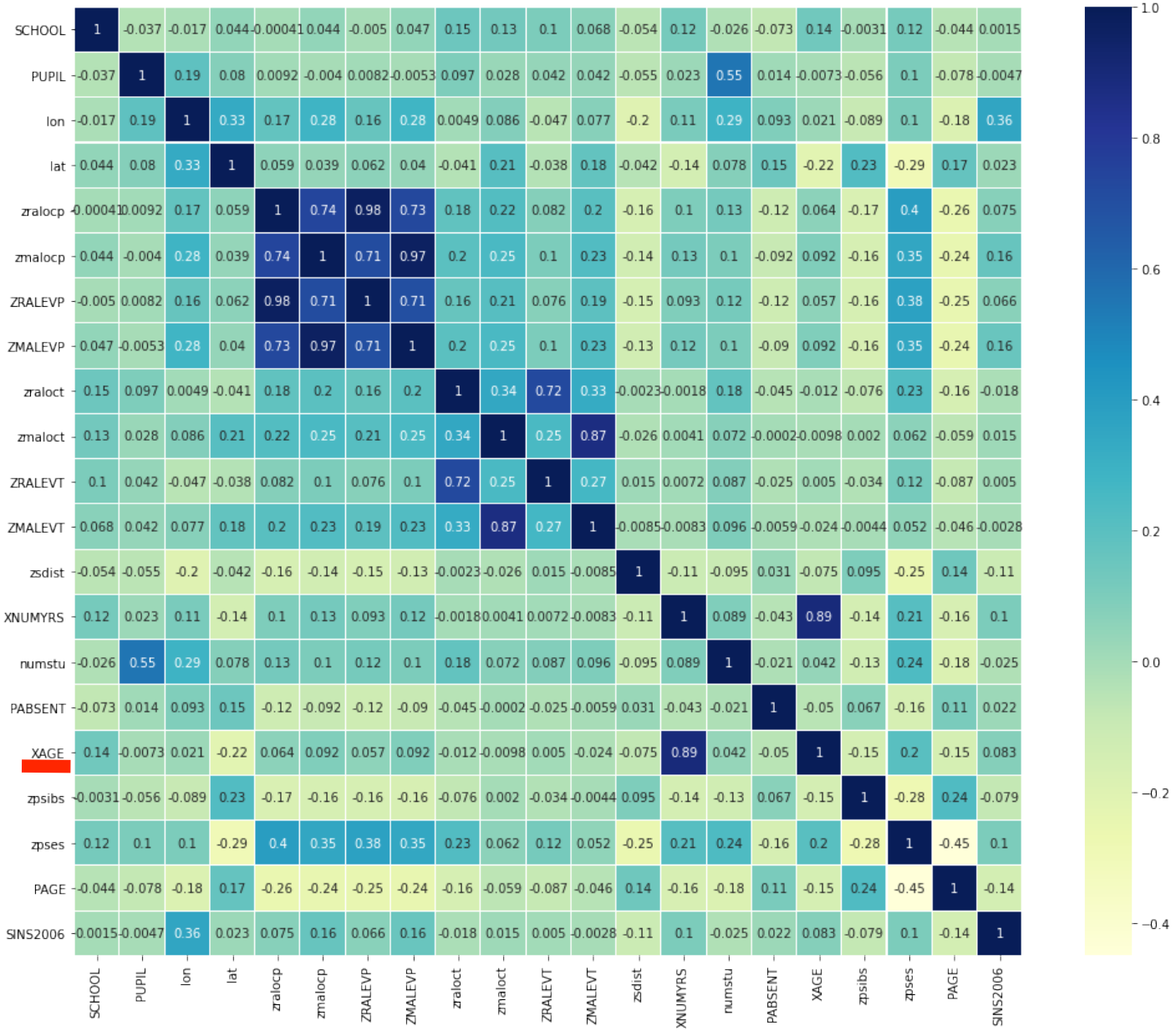
Pupils : 61 396
Columns : 2317
Countries : 15
Schools : 2779



Country
Region
District
School ID
School name
Pupil ID
Zralocp : Reading score's Pupil
Zmalocp: Mathematics score's Pupil
Zraloct : Reading competency level Teacher
Zmaloct : Math comptency level Teacher
XNUMYRS : Teacher Years of Teaching
numstu : Number of students at each school
PSEX : Pupil's Sex
PNURSERY : Pupil Preschool
PENGLISH : Pupil speaks English At Home
PTRAVEL : Travel To School
PTRAVEL2 : Means Of Travel To School
PMOTHER : Mother's Education
PFATHER : Father's Education
PLIGHT : Source Of Lighting
PABSENT : Days Absent***
PREPEAT : Repeated Grades
STYPE : School Type
SLOCAT : School Location
XSEX : Teacher's sex
XAGE : Teacher's Age
XQPERMNT : Teacher Employment Status
XQPROFES : Teacher Training
zpsibs : Pupil number of brothers and sisters
zpmealsc : Free school meals
zphmwkhl : Homework help
zpses : Pupil socioeconomic status
PAGE : Pupil's Age
SINS2006 : INSPECTION School
SPUPPR04 : Pupil Dropout
SPUPPR13 : Pupil Sexually Harrass Pupils
SPUPPR14 : Pupil Sexually Harrass Teachers

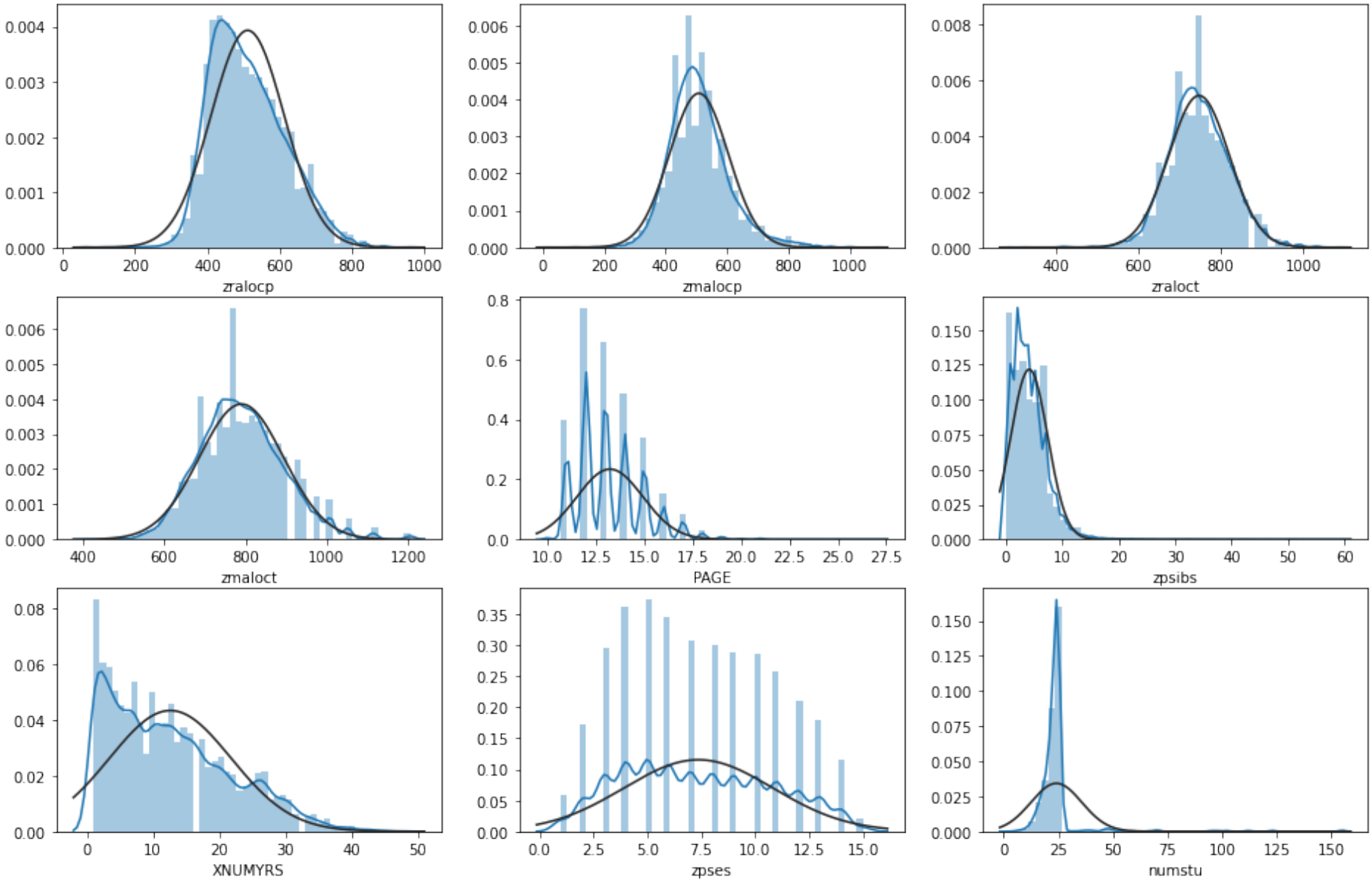
Data Exploration

Correlation Matrix



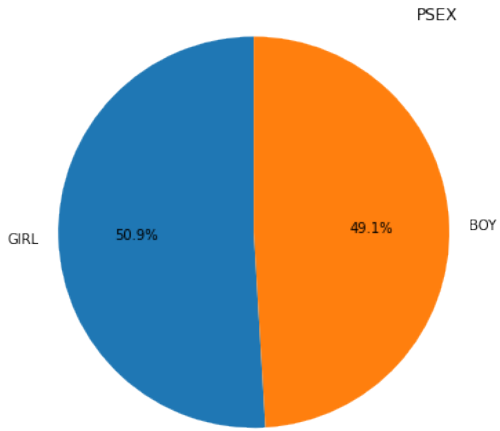
Data Visualization

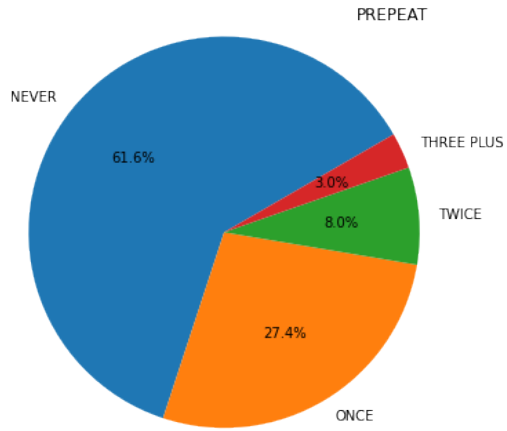
Distribution / Outliers

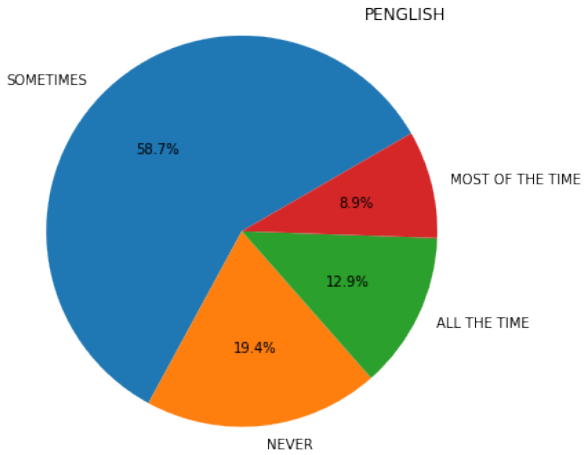


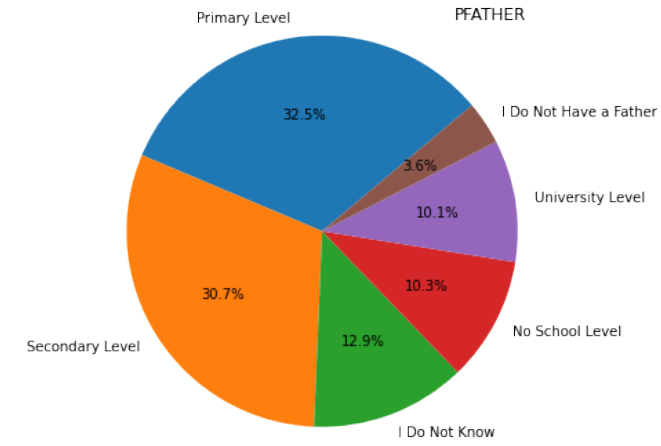
Data Visualization

Categorical variables repartitions









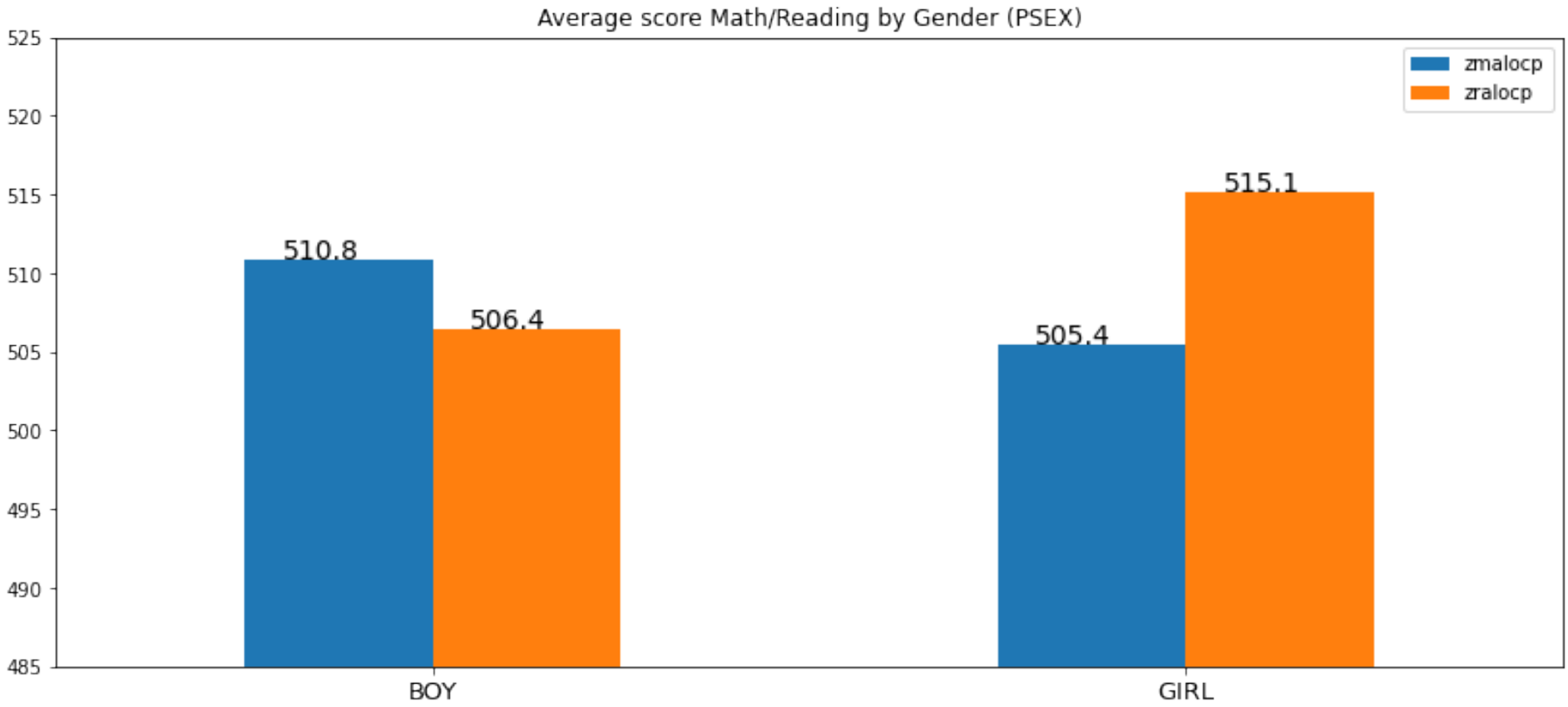
Data Preparation

Preprocessing

- Cleaning Data : dropping of collinear variables.
- Imputation of missing values.
- Dropping 2.65% of the missing values.
- Handling outliers ?
- Creating dummy variables
- Mean of the variable to be predicted : $score = (READ + MATH)/2$

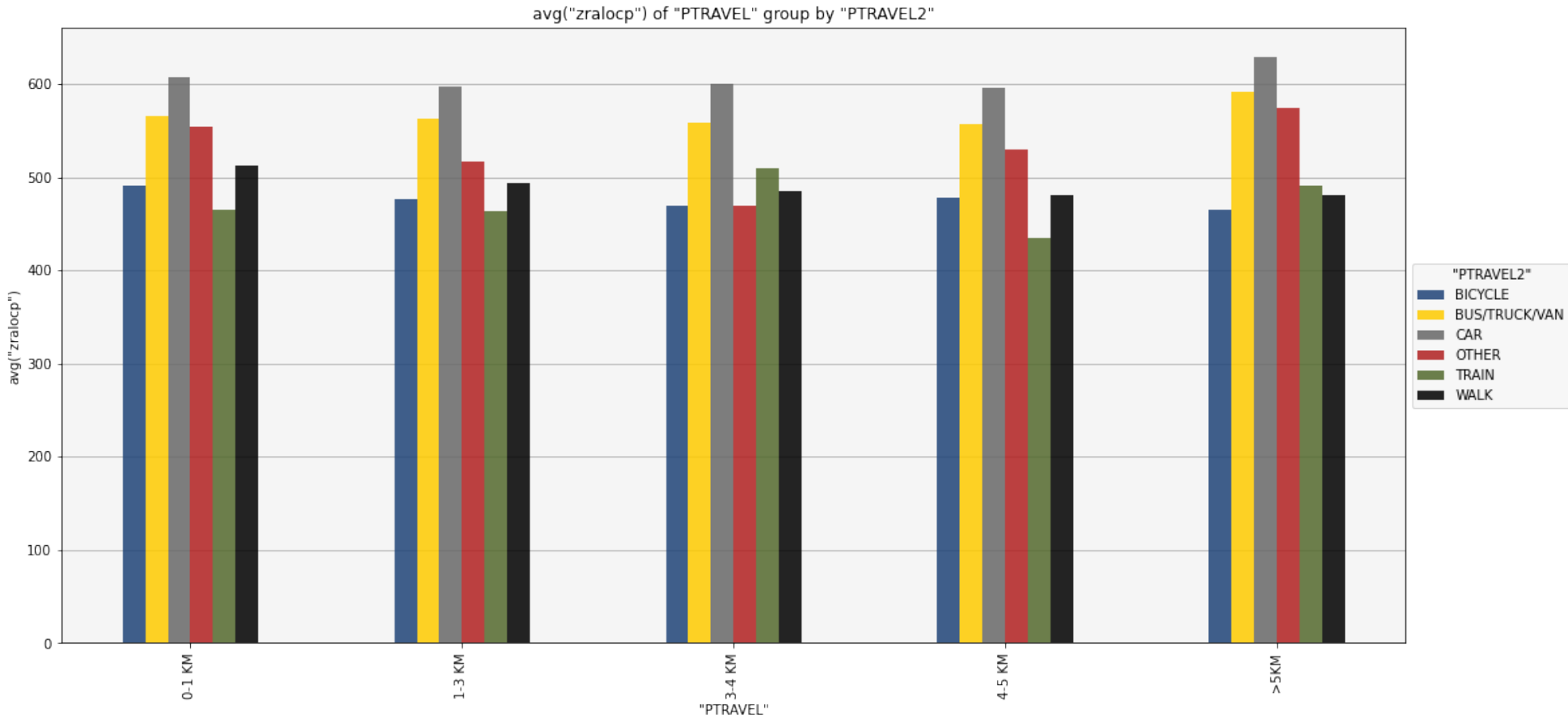
Relationship between target and features

First impressions : Pupil's sex



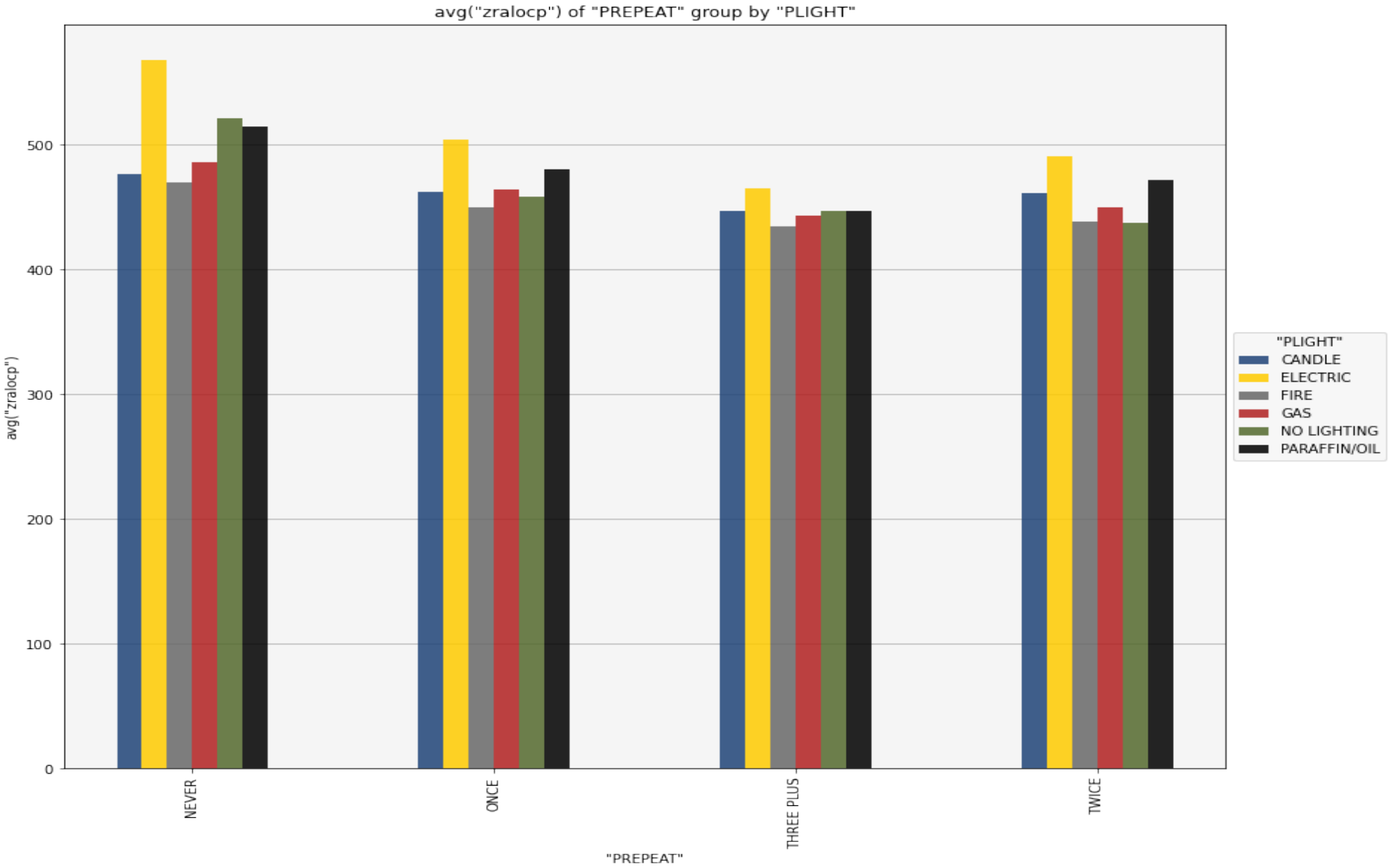
Relationship between target and features

First impressions : Pupil's means of travel to school



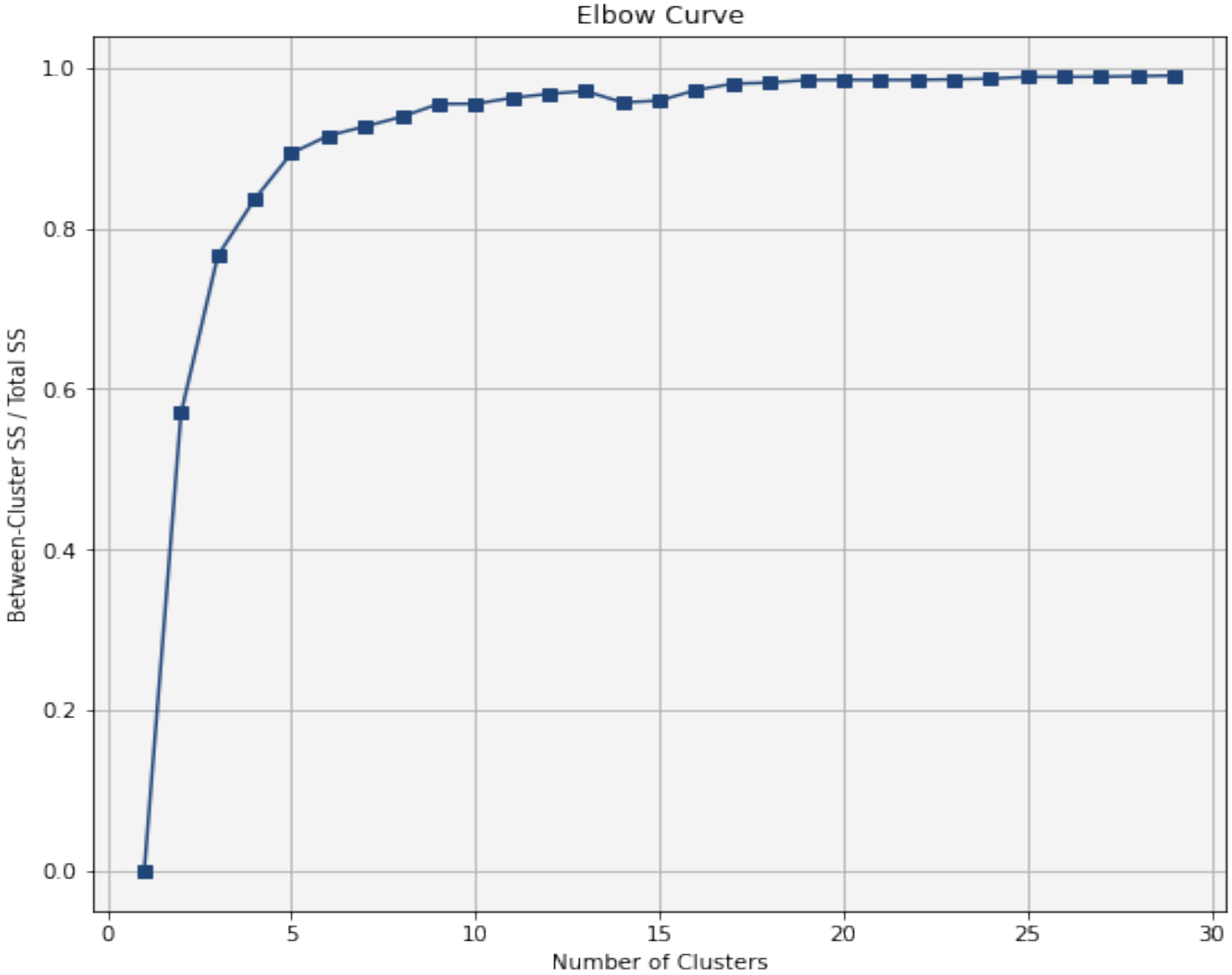
Relationship between target and features

First impressions : Source of lighting



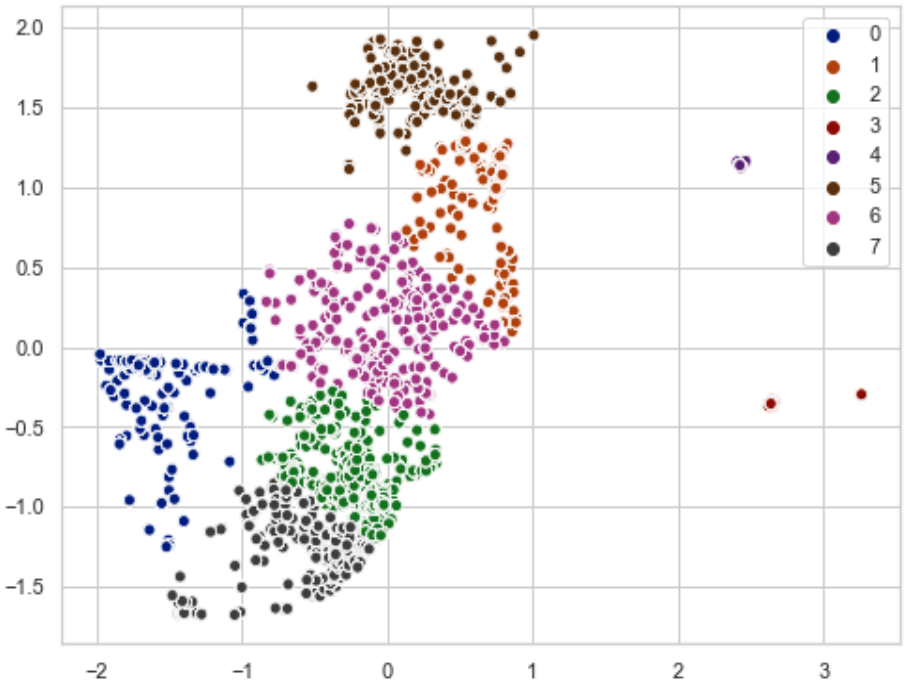
Machine Learning Unsupervised

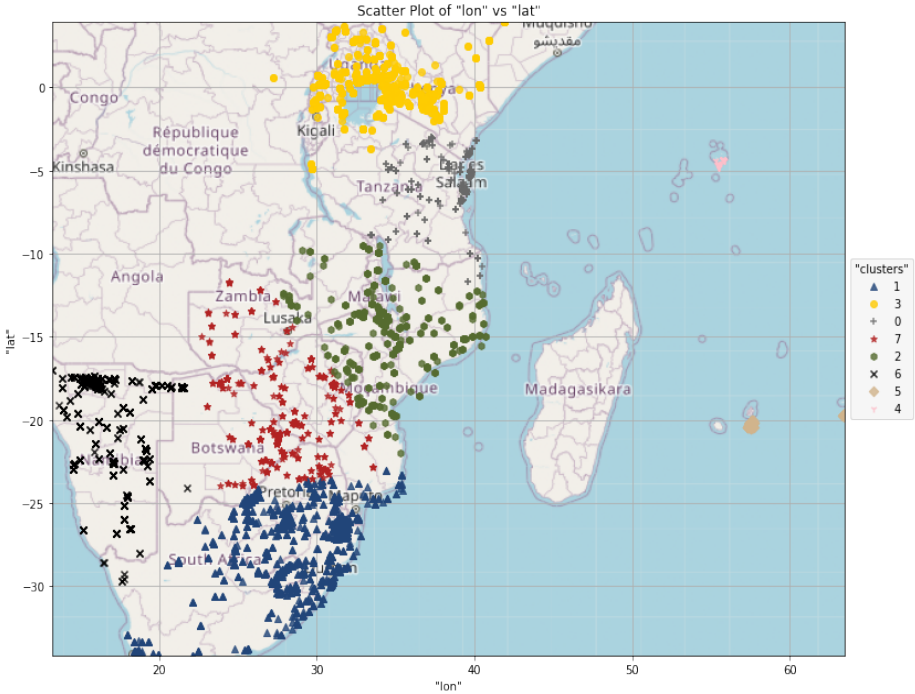
Finding Clusters using lat/long : Elbow method



Machine Learning Unsupervised

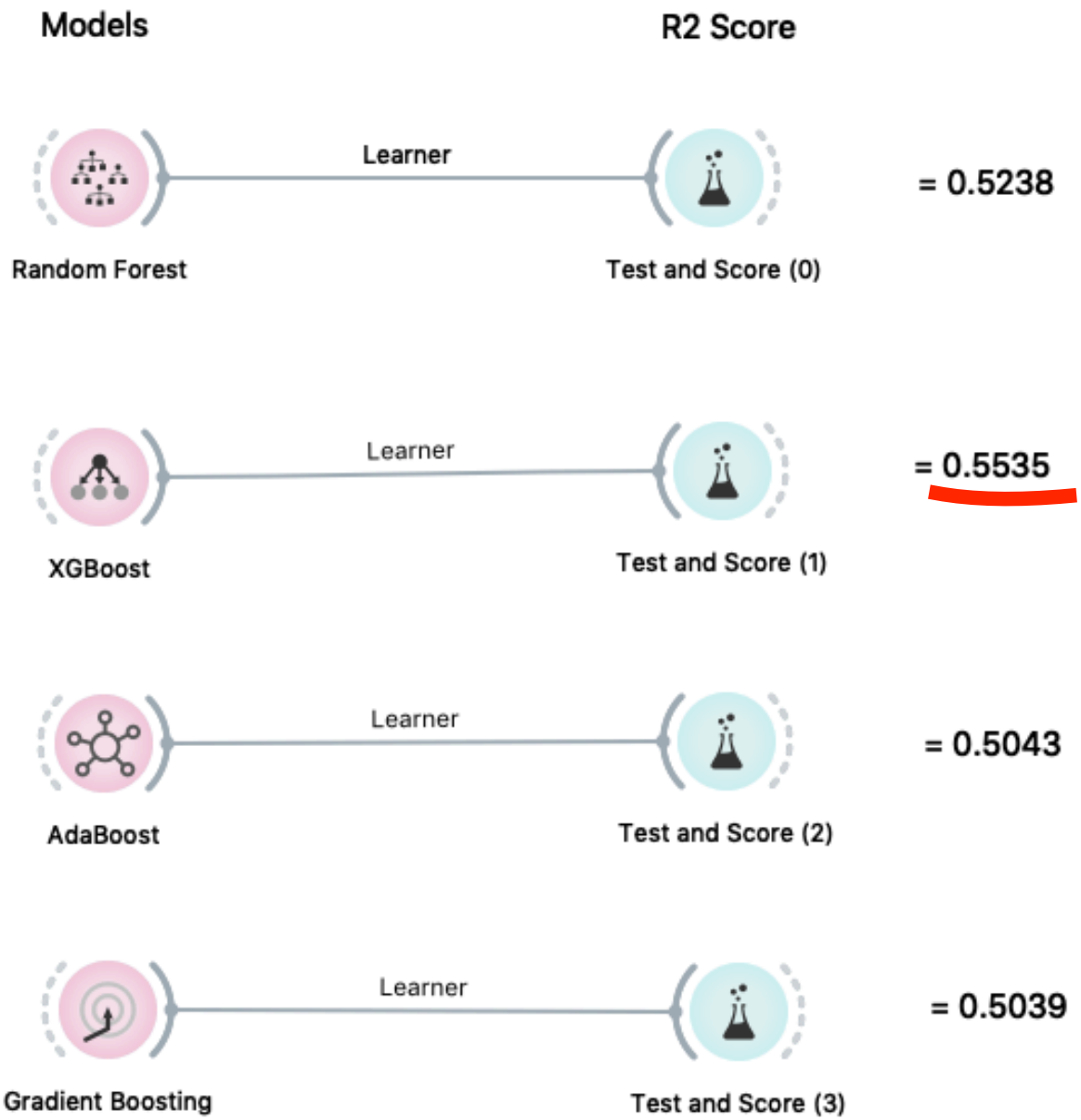
KMeans model





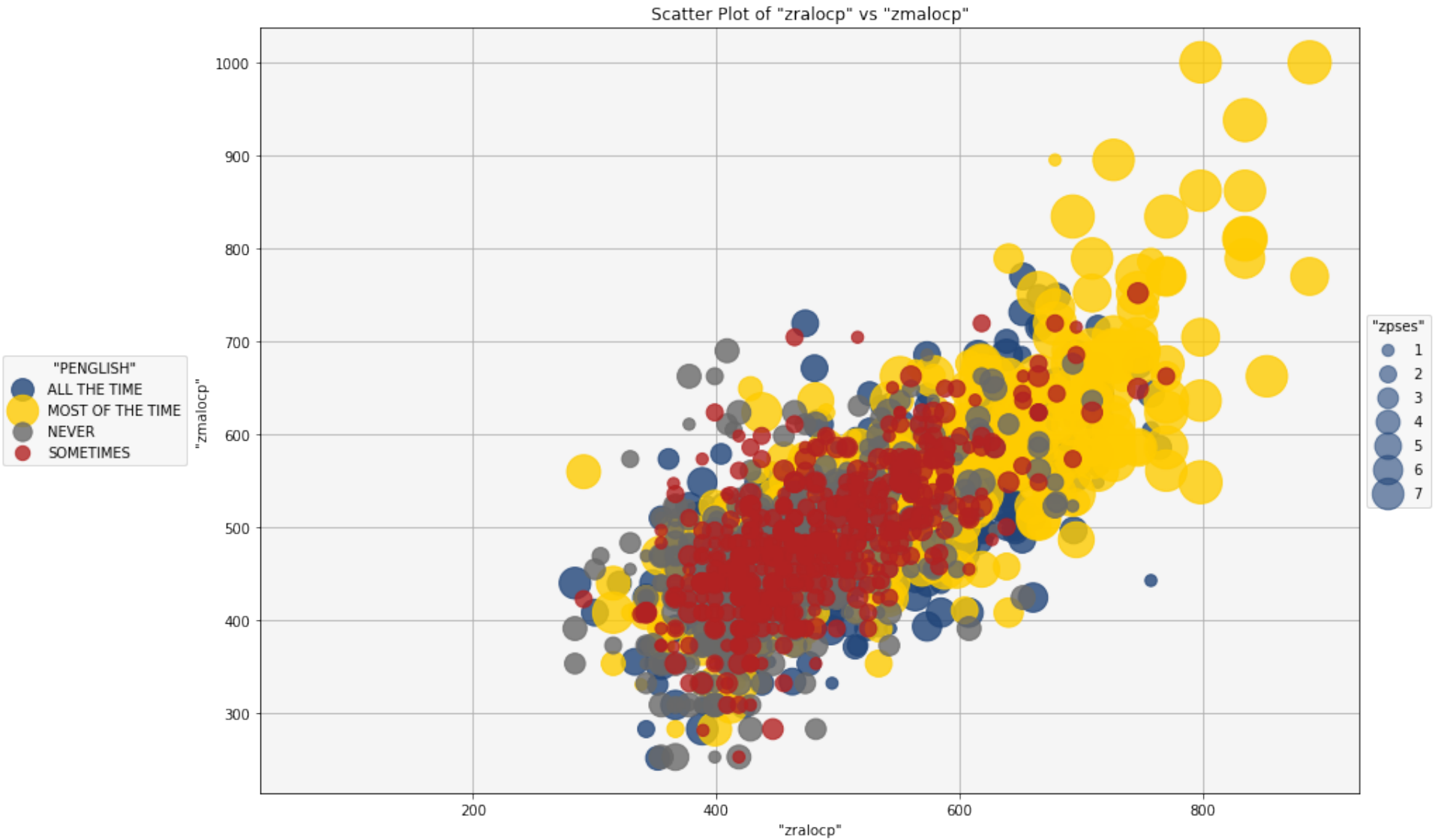
Machine Learning Supervised

Model performance metrics



Machine Learning Supervised

Typical profile of the more successful student



Conclusion, Recommandation & Perspective