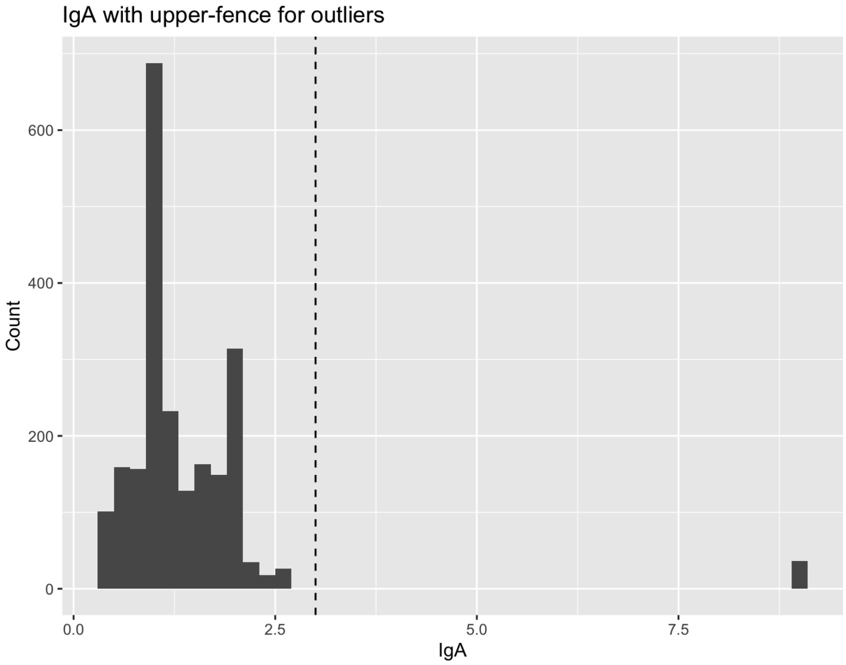
Predicting Celiac Disease Diagnosis from Clinical and Lab data

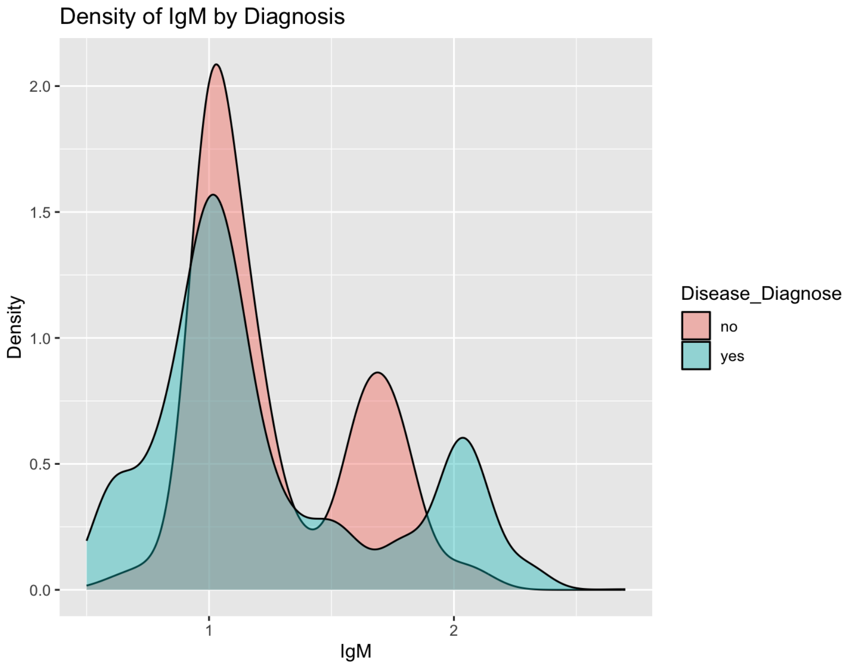
Descriptive Analysis

The dataset contained **2,206** sample points with a roughly even gender distribution (female = 1,084; male = 1,122). Approximately 93.5% of the sample was diagnosed with celiac disease.

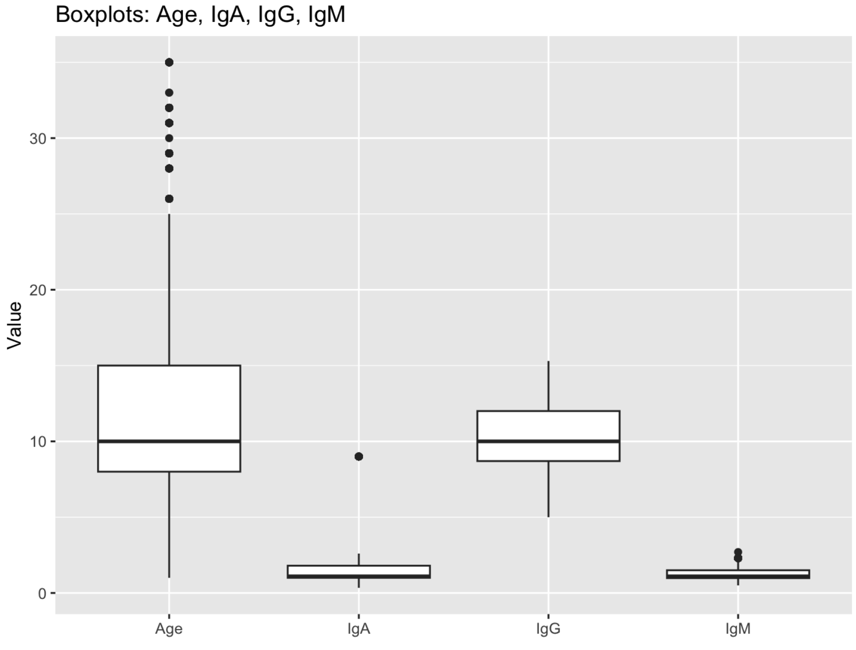
About the data:
Continuous variables:

- Age: ranged 1-35 years; mean 12.8, median 10 with a slight skew towards children
- IgA (immunoglobulin A) : ranged 0.34-9.0 g/L; mean 1.43, with a single outlier (9.0)
- IgG (immunoglobulin G): ranged 5.0-13.3 g/L; mean 10.1 g/L
- IgM (immunoglobulin M): ranged 05-2.7 g/L; mean 1.24 g/L





Density of IgM by Diagnosis



Boxplots: Age, IgA, IgG, IgM

Comparative Analysis

I started by testing whether the dataset reflected a slight correlation with females having a higher proportion of celiac disease.

	No Celiac	Yes Celiac
Female	155	929
Male	208	914

Proportions: 86% of females and 81% of males were diagnosed with Celiac disease
Chi square test: $\chi^2(1) = 6.90$ giving a $p = 0.0086$ which is statistically significant
Odds ratios: males had 0.73 times the odds of being diagnosed (95% CI: 0.58-0.92).

Predictive Modeling

I chose to fit a logistics regression model to predict celiac diagnosis from demographics, symptoms, and immunological markers. I chose to drop "Marsh" as a variable because it is redundant and would skew my model's accuracy upwards.

Significant predictors:

- IgA (OR ≈ 3.3 per unit, $p < 0.001$)
- IgM (OR ≈ 2.8 per unit, $p = 0.0007$)
- Diabetes Type 1 (OR ≈ 119, $p < 2e-16$)
- Diabetes Type 2 (OR ≈ 100, $p = 1.8e-11$)
- Abdominal pain (yes) (OR ≈ 6.0, $p < 1e-7$)
- Short stature (PSS) (OR ≈ 2.8 vs DSS, $p = 0.003$)
- Diarrhoea type: "Inflammatory" and "Watery" were less predictive than "Fatty" ($p < 0.05$).

Non-significant predictors:

- age, gender, IgG, short stature, sticky stool
- weight loss was almost significant with a p value of 0.08

Model fit:

- Null deviance = 1390.26 which gave a residual deviance of 567.69
- AIC = 597.7

Takeaways

Immune markers (IgA, IgM) and comorbid diabetes were the strongest predictors of diagnosis. Abdominal pain and growth impairment (short stature) added significant predictive value but themselves were not enough. Data supports modest gender differences with statistically significance in comparative tests, though gender was not an independent predictor once other factors were controlled. Overall, The model is both statistically strong as it has a large deviance reduction and clinically interpretable since the OR values align with known risk factors.