What other factors can be used to predict cereal yield?

Wen Fung Leong

Motivation

- Find a linear model to establish the relationship between specific factors and cereal yield for each country
- Time frame years 1991 – 2015 (25 years)

Features and Target (1)

Target	Ckg	Cereal yield (kg per hectare)

Categories	**Features**	**Short Description**
Nature	A	Agricultural arable land (% of land area)
	Cpl	Land under cereal production (hectares)
	Ravg	Average precipitation in mm
	Tavg	Mean air temperature in ^{o}C
Population	P	Total population (in millions)
	R	Rural population (% of total population)
Technology	F	Fertilizer consumption (kilograms per hectare of arable land)
	M	Agricultural machinery, tractors per 100 sq. km of arable land

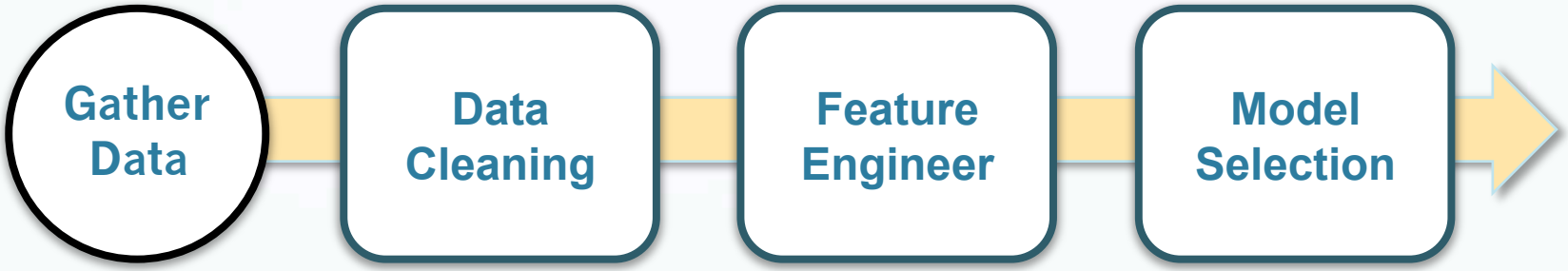
Features and Target (2)

Categories	Features	Short Description
Economy	V	Agriculture, forestry, and fishing, value added (% of GDP)
	G	GDP (current US$)
	Gc	GDP per capita (current US$)
	E	Employment in agriculture (% of total employment)
Cost	DI	Pump price for diesel fuel (US$ per liter)
	GI	Pump price for gasoline (US$ per liter)

Total features = 14

Correction on the number of feature.
Target was included by mistake.

Methodology

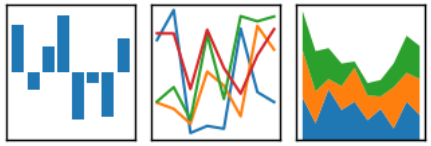


seaborn

Food and Agriculture Organization of the United Nations

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



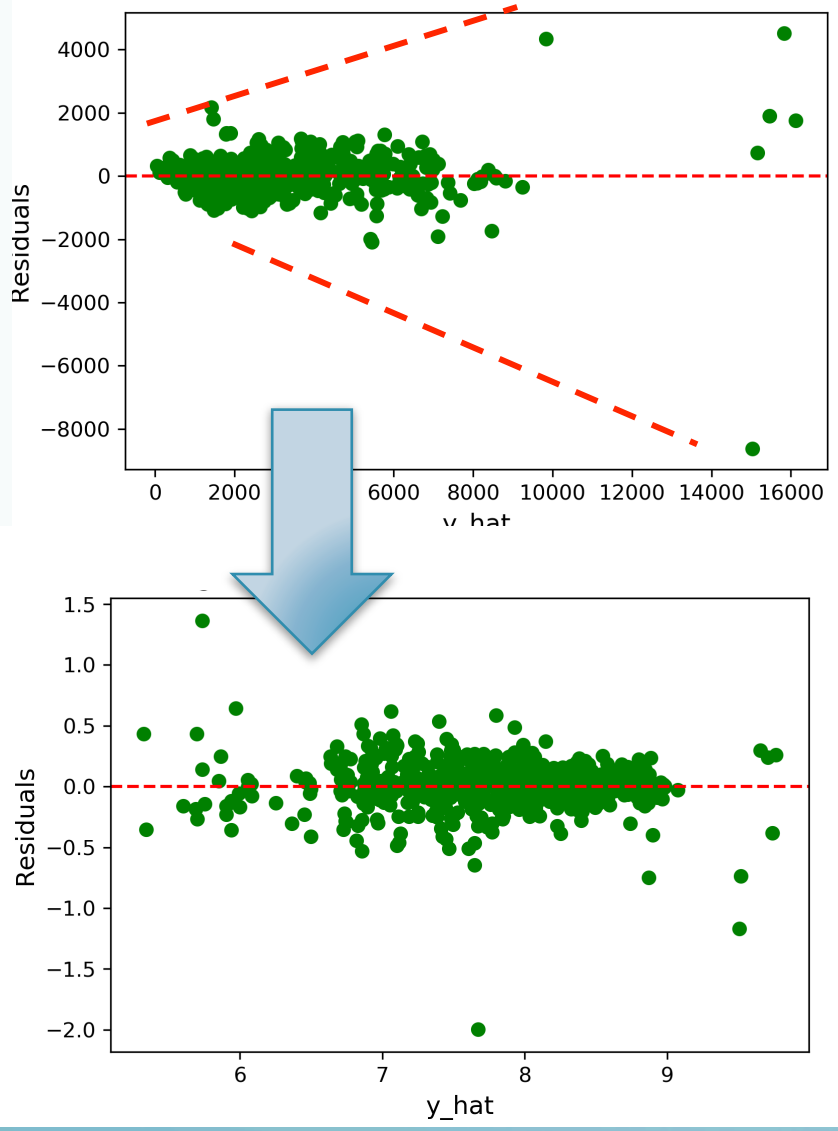
WORLD BANK GROUP

Data Cleaning

- 118 countries with data (removed 88)
- Imputation to handle missing data
  - Linear Regression
  - Substitution (minimum or 0)
- Final total # of features = 131
  - Excluded GDP
  - 13 + 118 categorical variables

Correction: Total # countries – 206
Remainder is 118 after removing 88 countries

Feature Engineer



Heteroscedasticity

$R^2 = 0.917$
Adj $R^2 = 0.893$
RMSE = 660.88

After log transforming the cereal yield

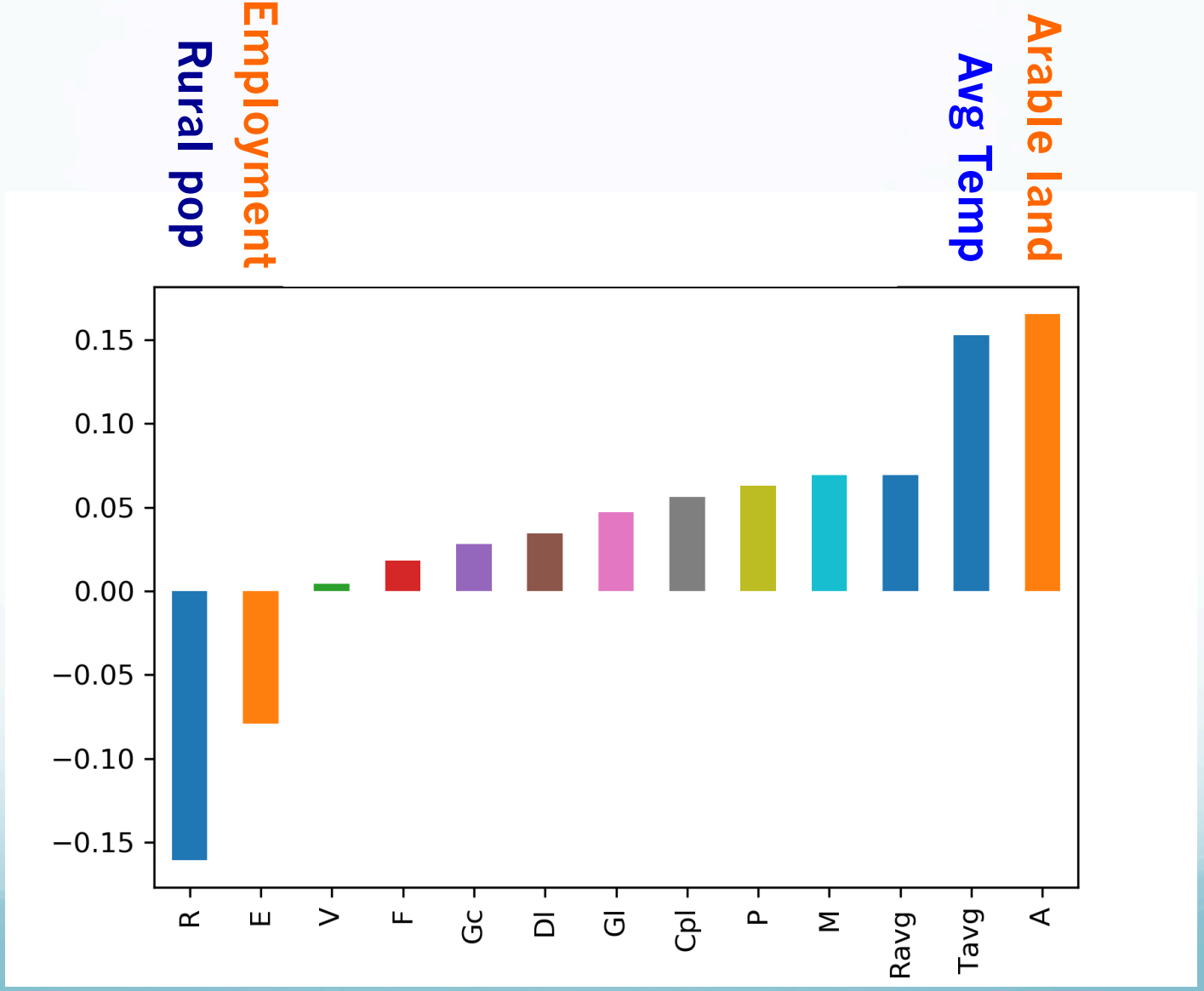
$R^2 = 0.921$
Adj $R^2 = 0.899$
RMSE = 0.210

Model Selection

- Cross Validation: 10 folds

Linear Models	Metrics
Linear Regression	$R^2 = 0.913 \pm 0.02$
Lasso (alpha = 0.0033)	$R^2 = 0.910 \pm 0.022$
Ridge (alpha = 3.1578)	$R^2 = 0.573 \pm 0.021$

Regression Coefficients



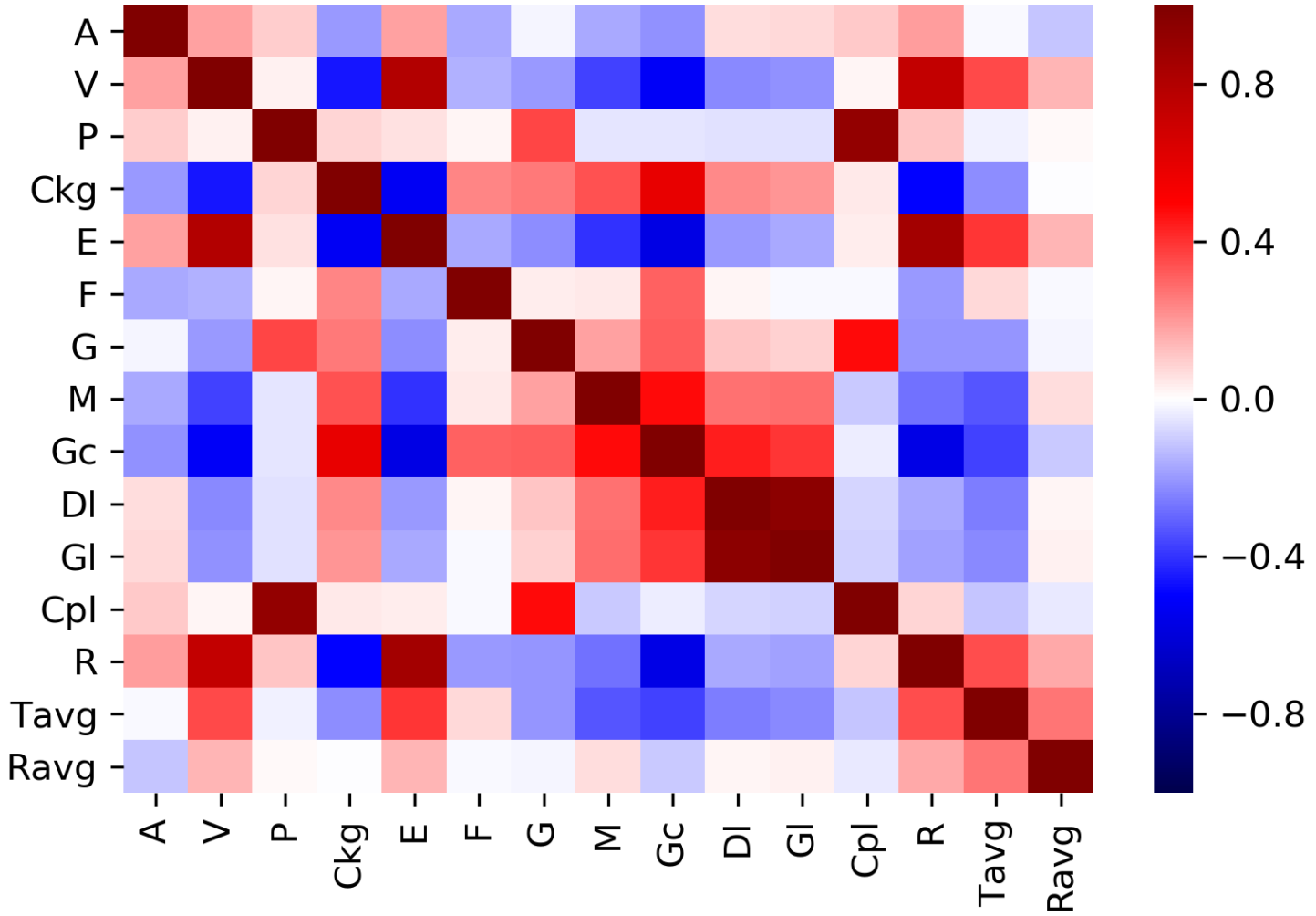
Future Work

- Compare the model's prediction with 2016 – 2018 cereal yield data
- Add interaction terms on the regression model
- Expand time frame (before year 1991)
- Consider other variables (Infrastructure, Agricultural R & D spending)


Thank You

Appendix

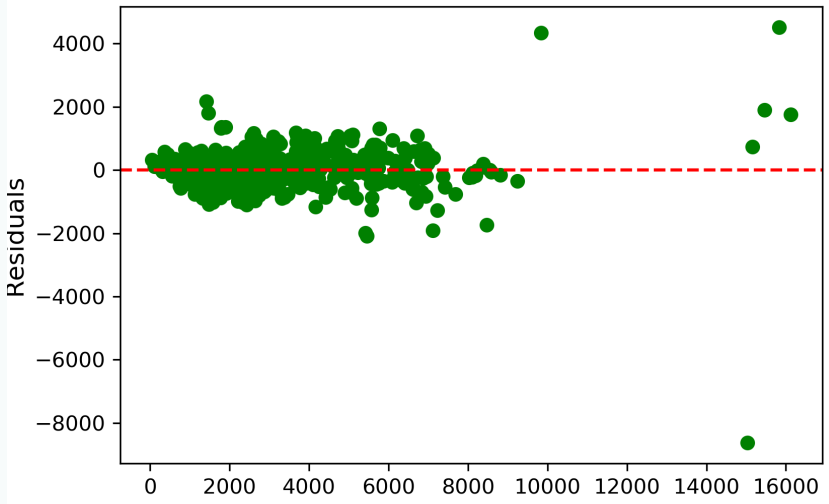
Data Cleaning and EDA

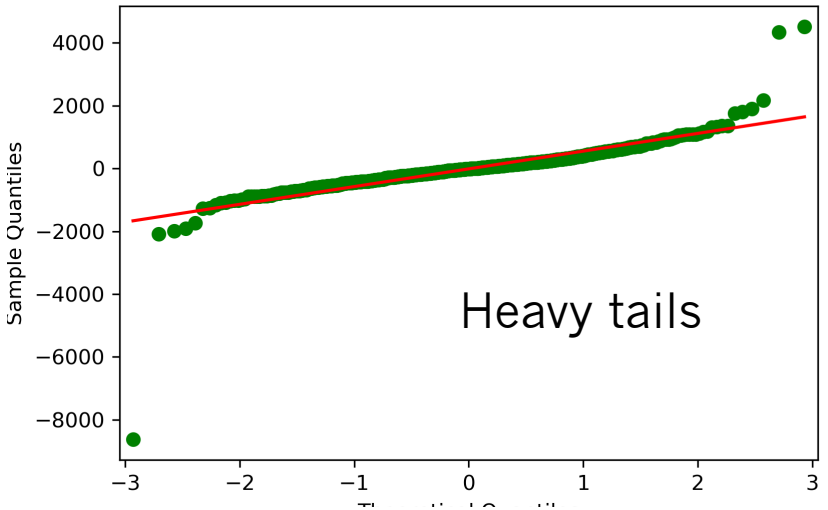


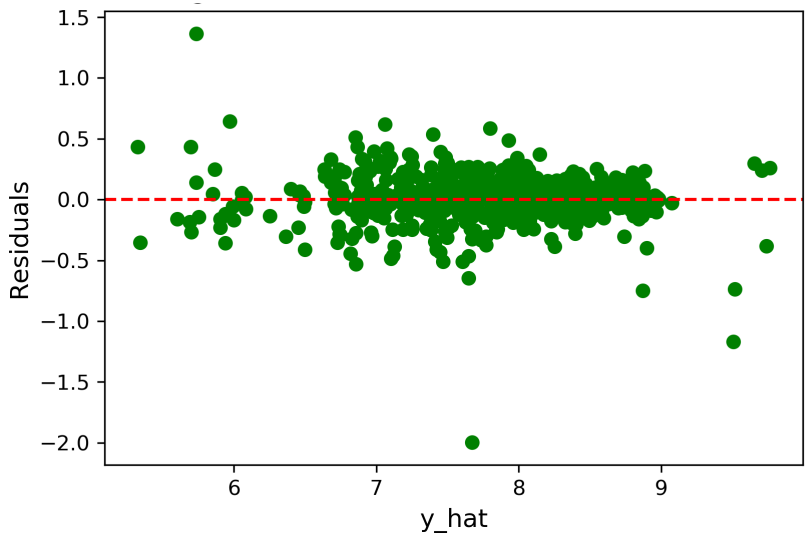
Data Cleaning and EDA

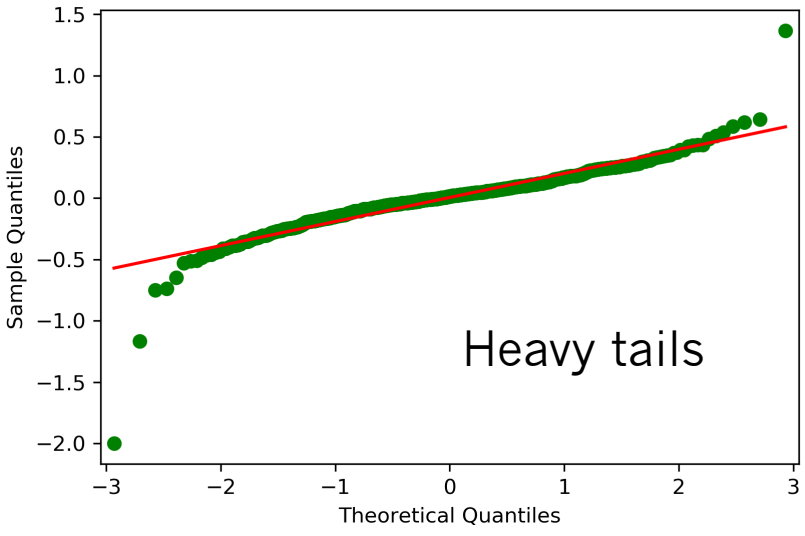
Imputation	Feature
Linear Regression	Dl, Gl, M, P, R
Substitution (with 0)	F
Subtitution (with min value)	V, Ckg, G, Gc, Cpl

After Feature Engineer (1)

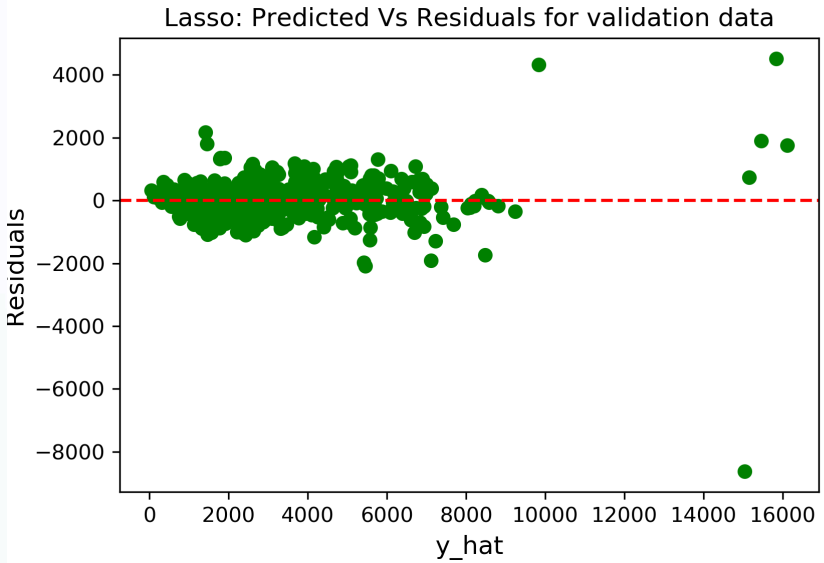


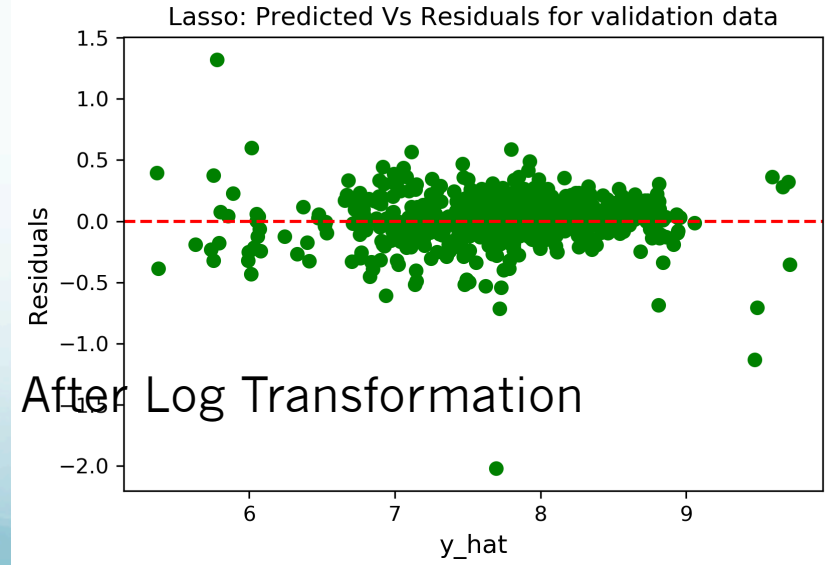






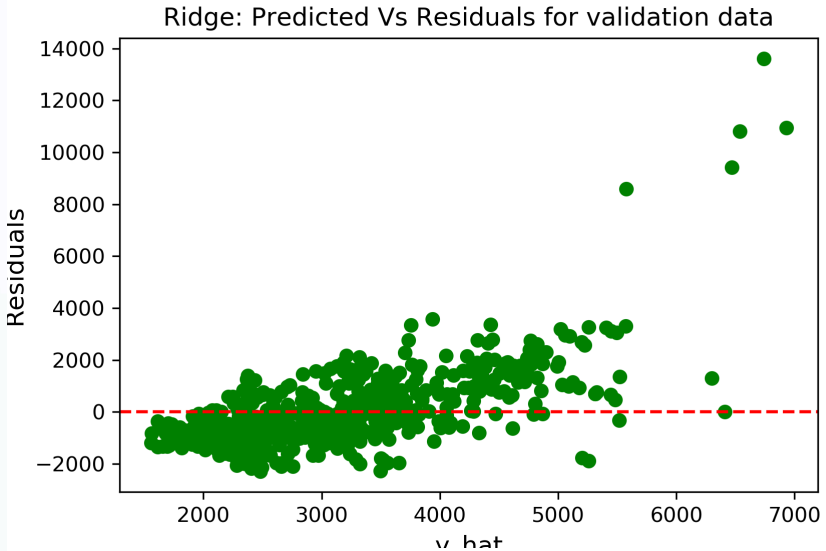
After Feature Engineer - Lasso

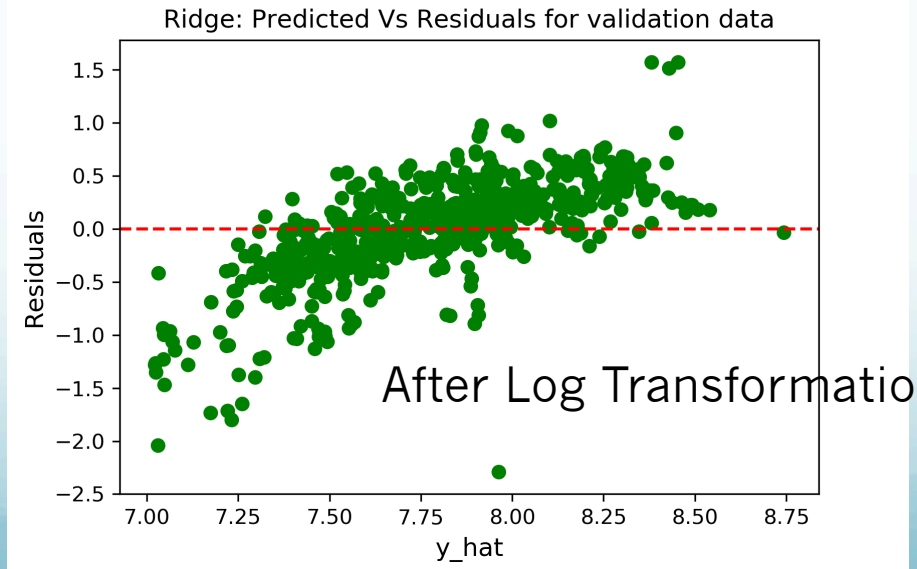




After Log Transformation

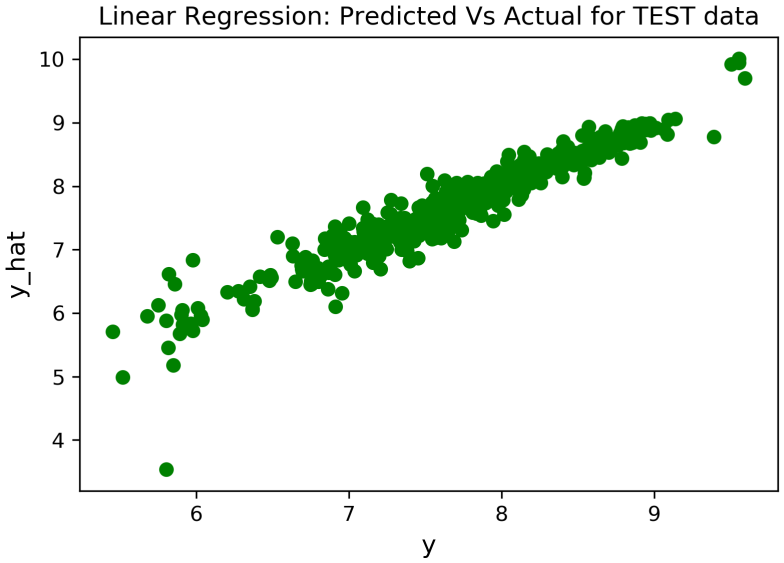
After Feature Engineer - Ridge

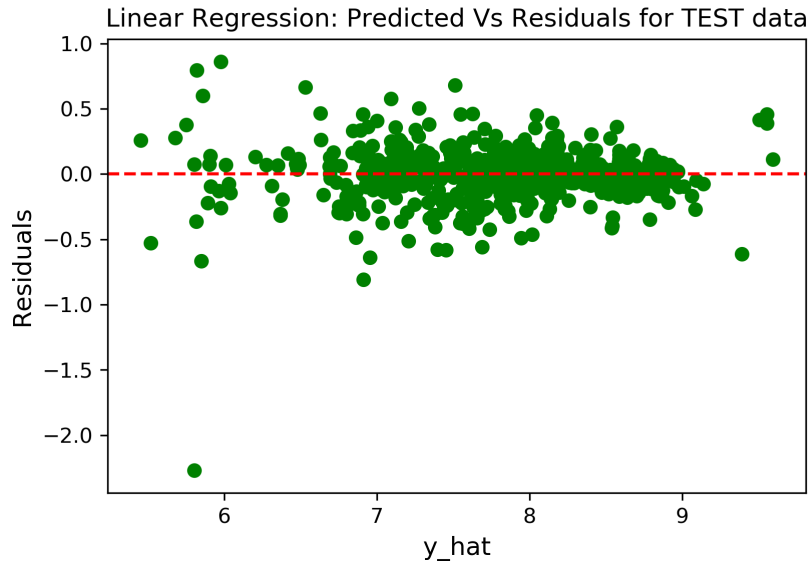




After Log Transformation

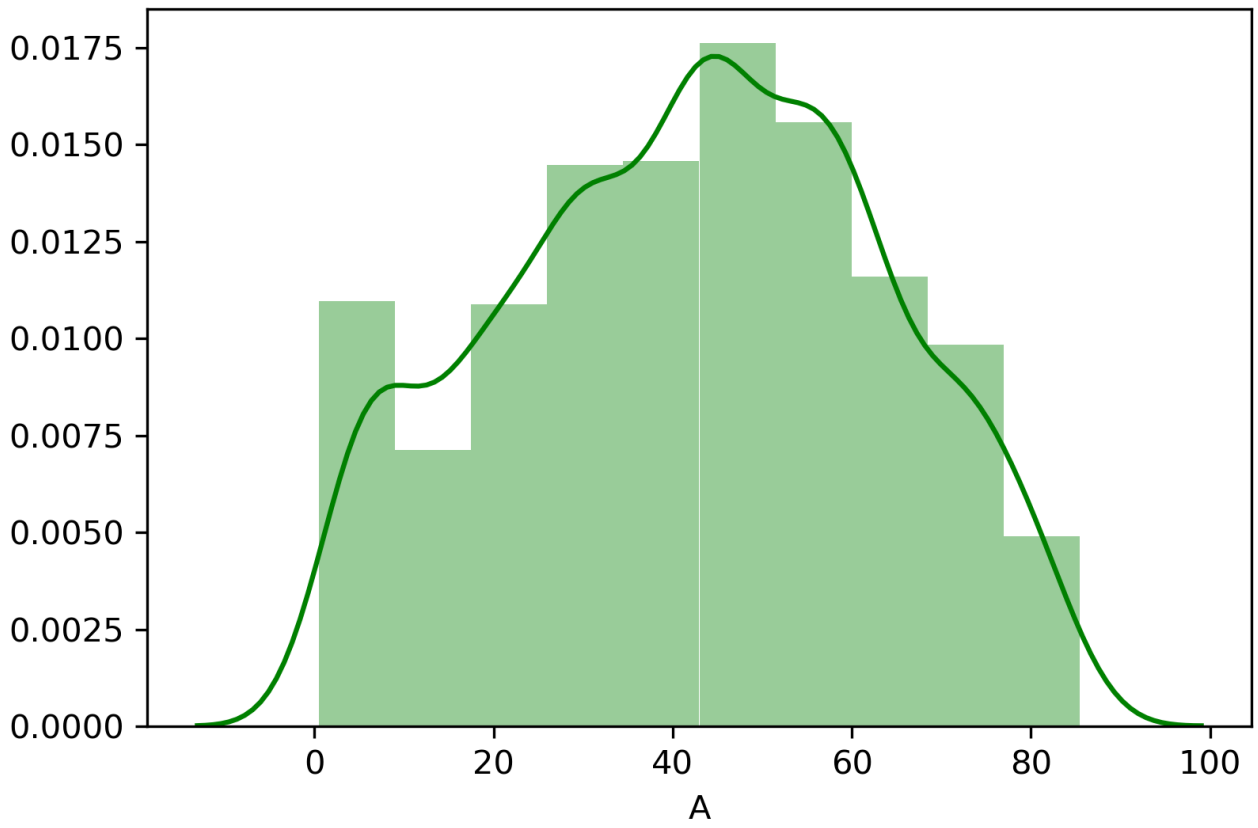
Final run on the test set



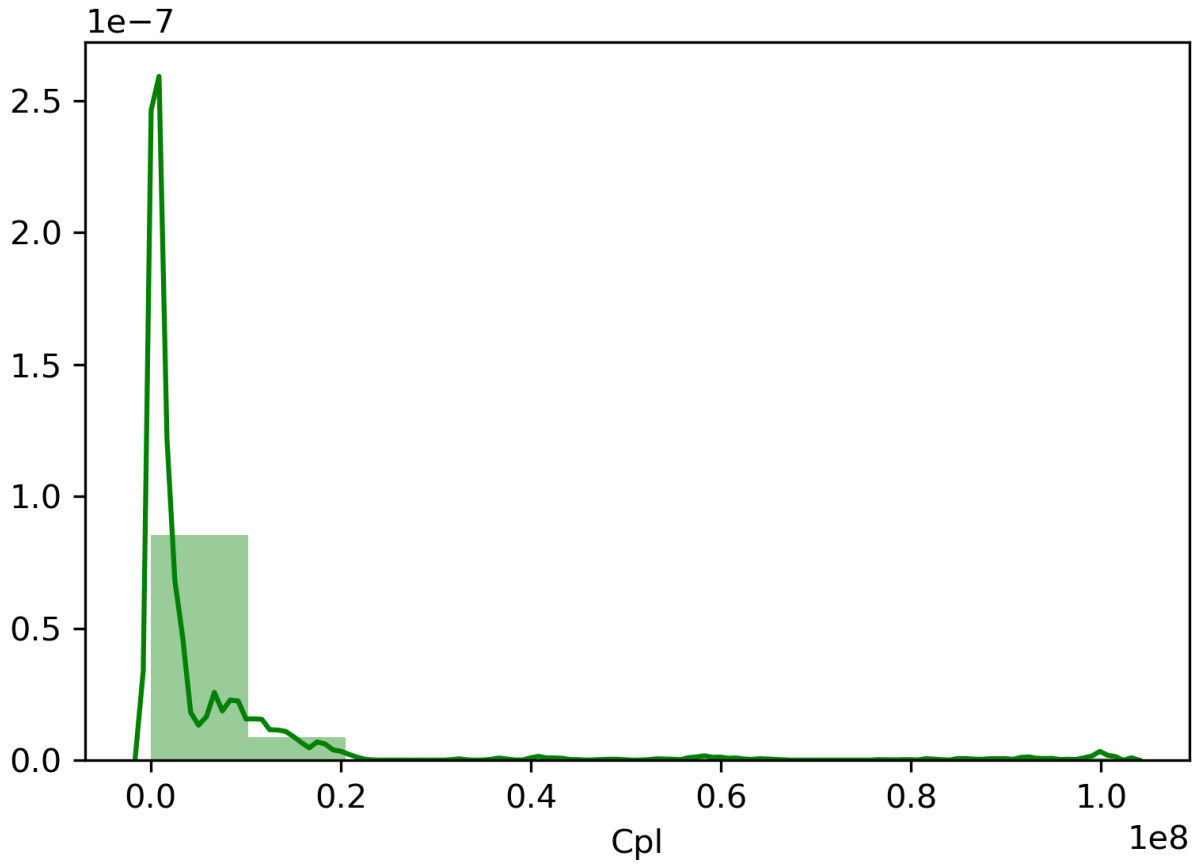


$R^2 = 0.922$
Adj $R^2 = 0.899$
RMSE = 0.214

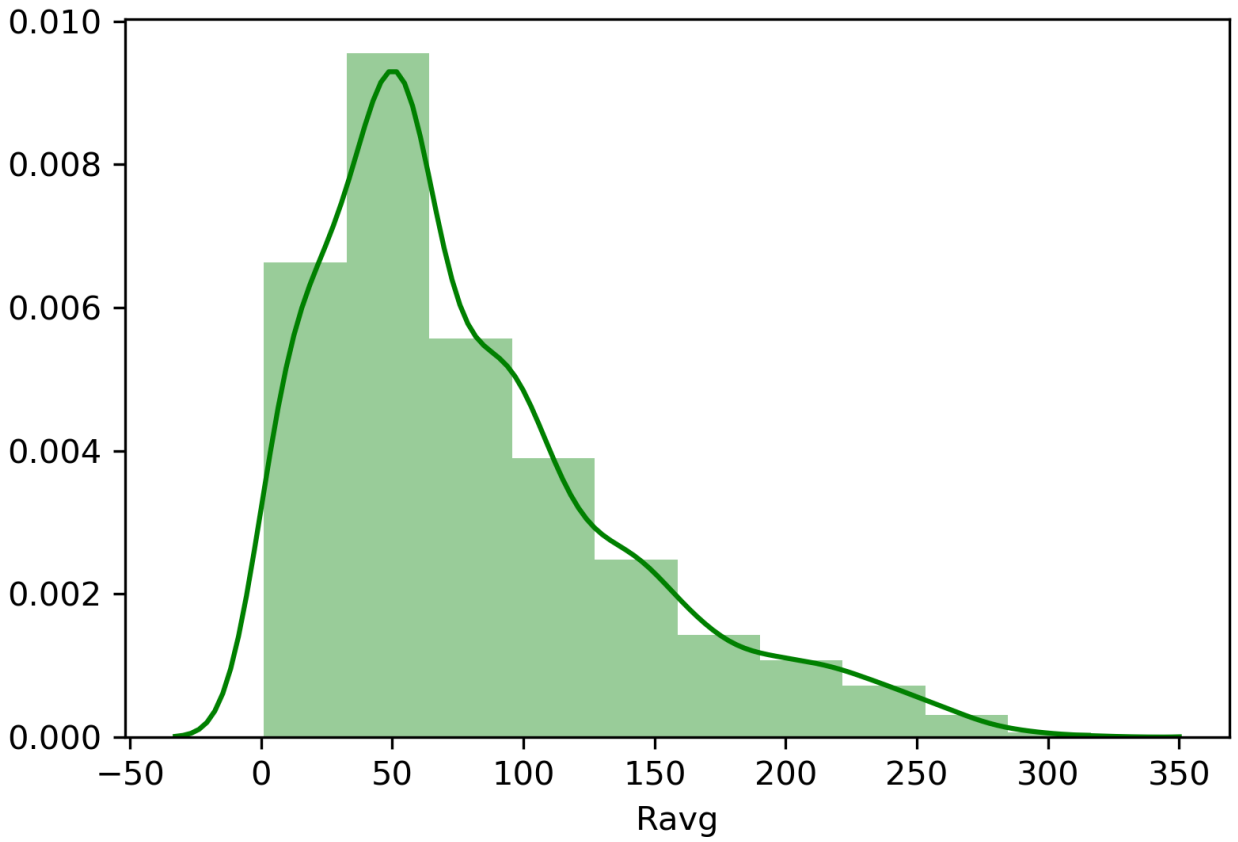
Agricultural land (% of land area)



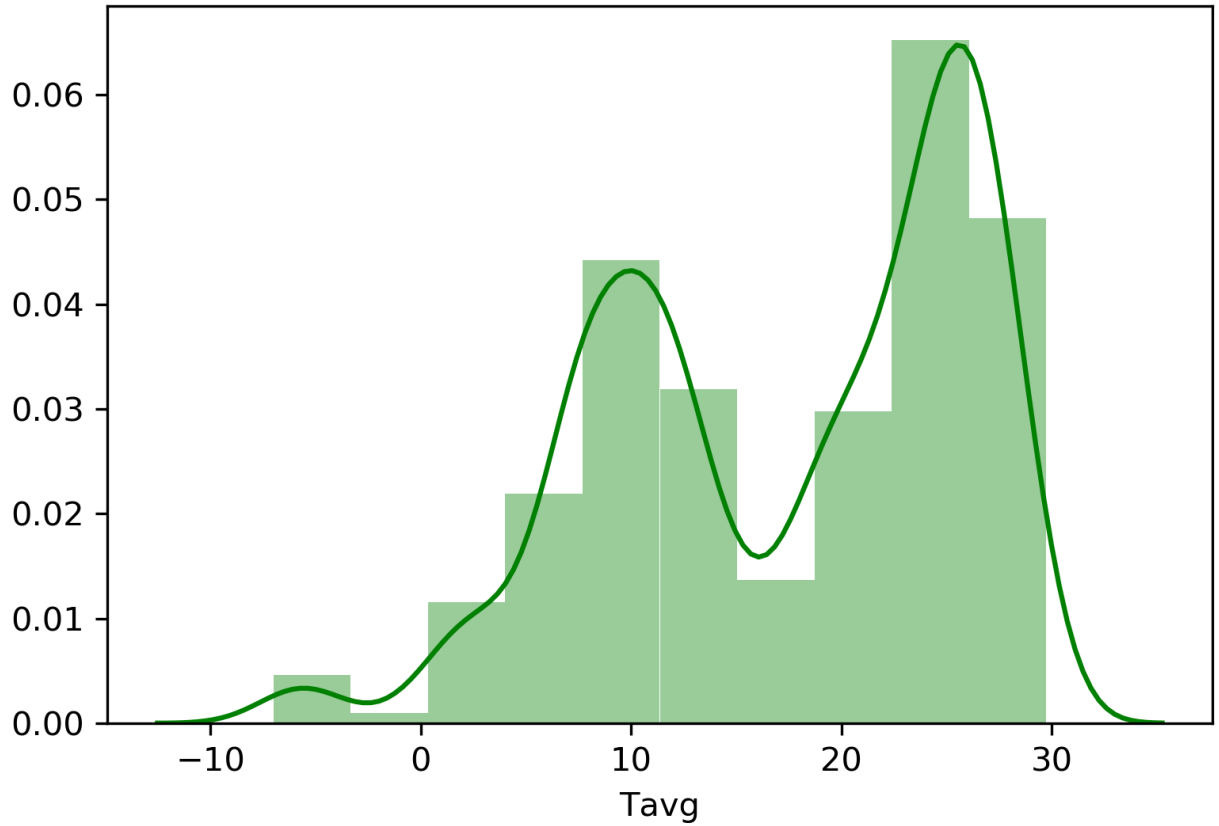
Land under cereal production (hectares)



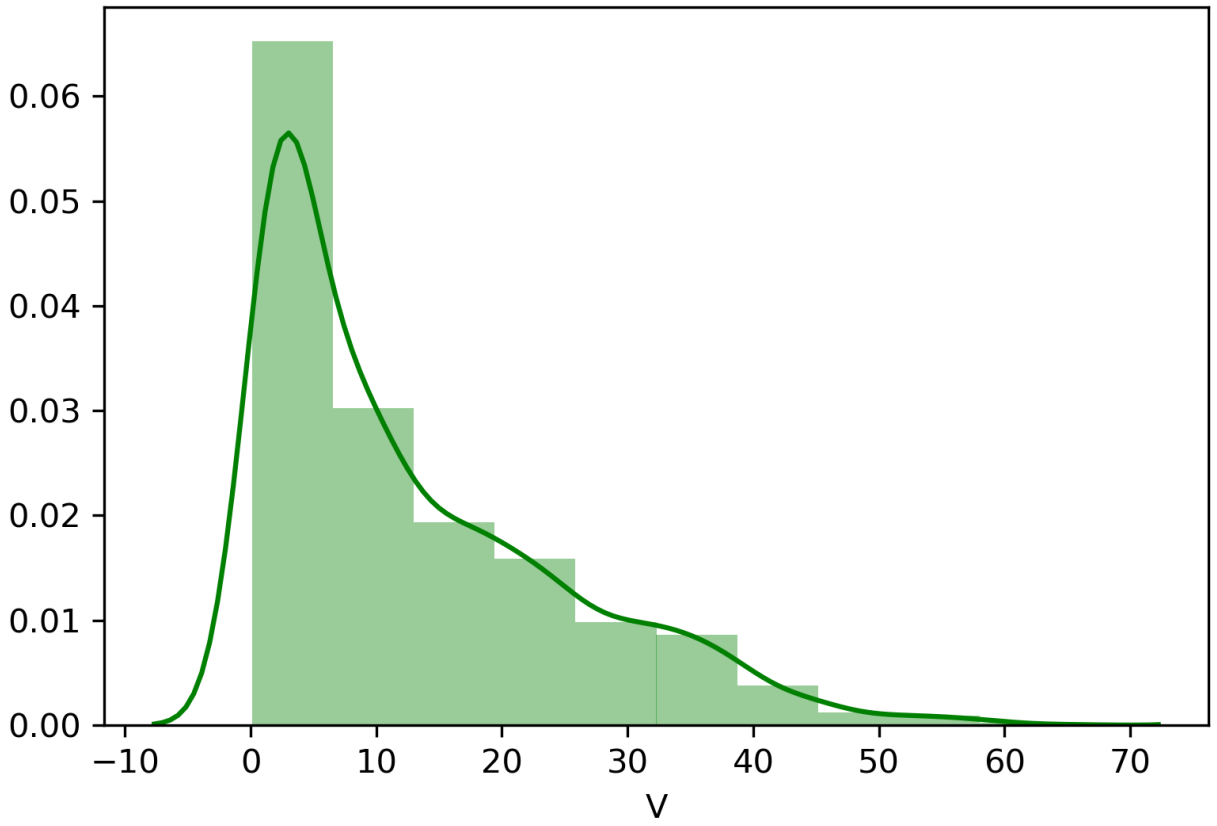
Average precipitation



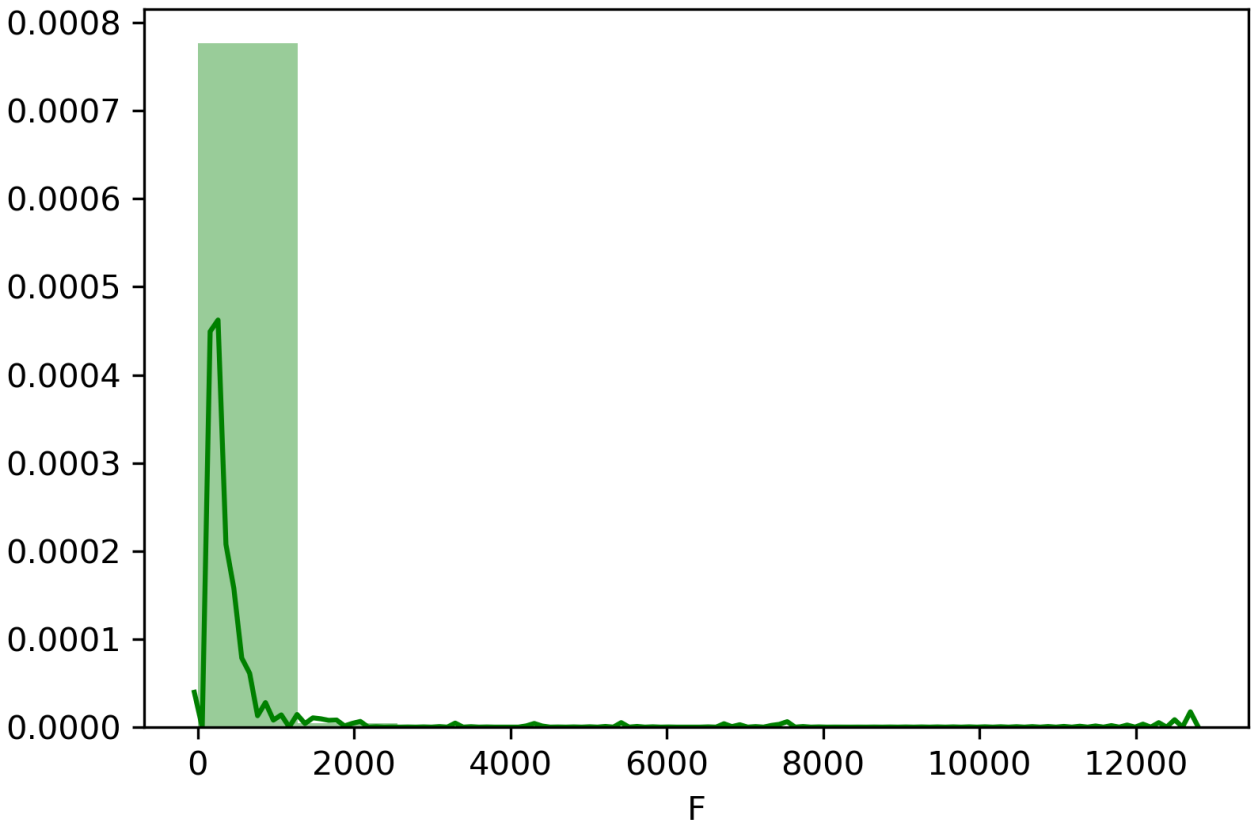
Average temperature



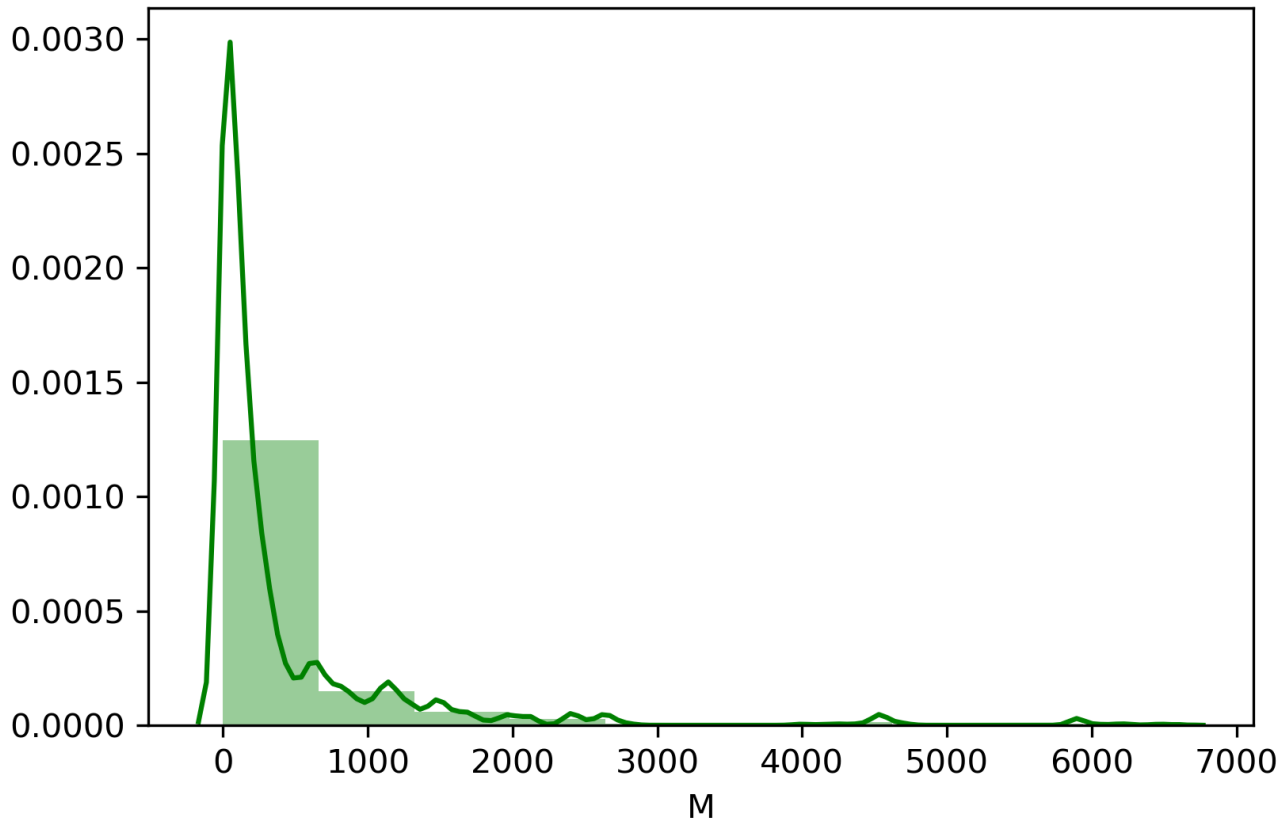
Agriculture, forestry, and fishing, value added (% of GDP)



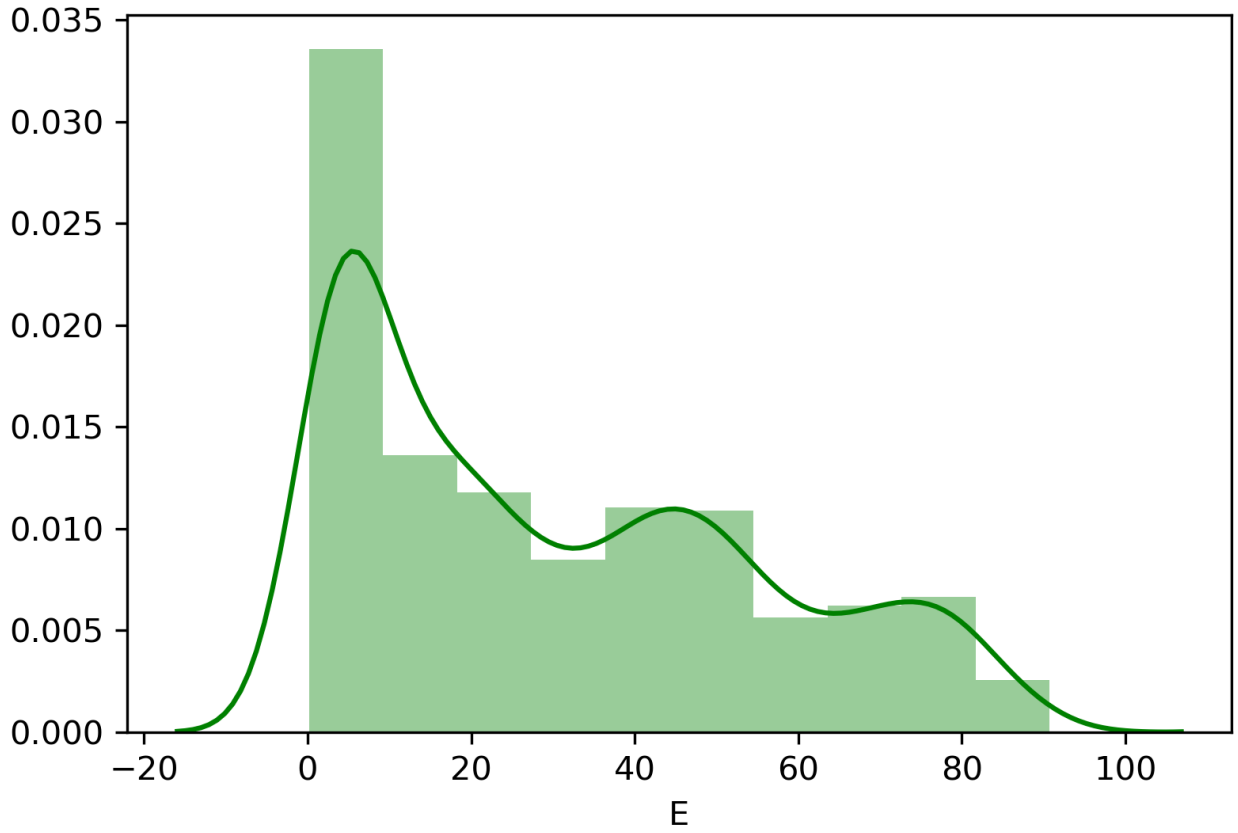
Fertilizer consumption (kilograms per hectare of arable land)



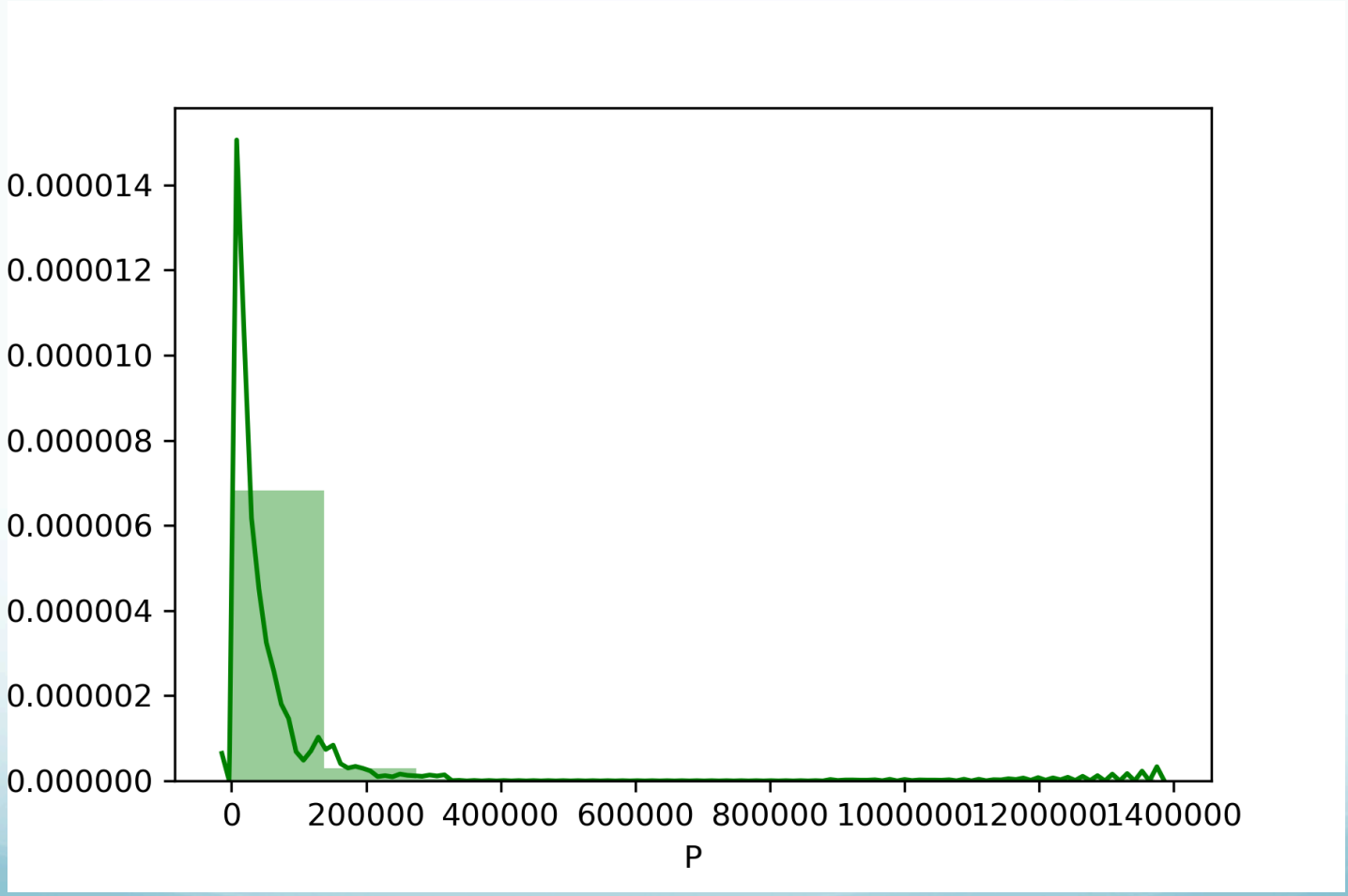
Agricultural machinery, tractors per 100 sq. km of arable land)



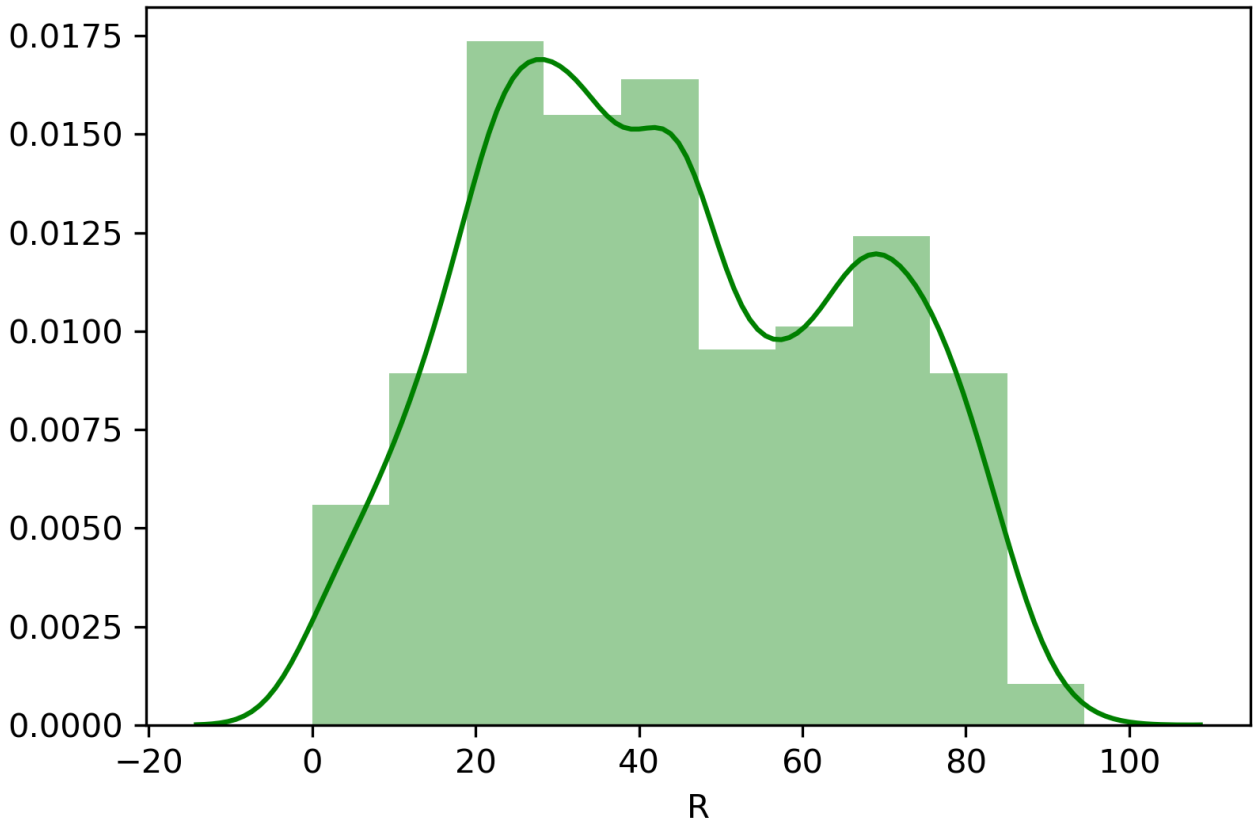
Employment in agriculture (% of total employment)



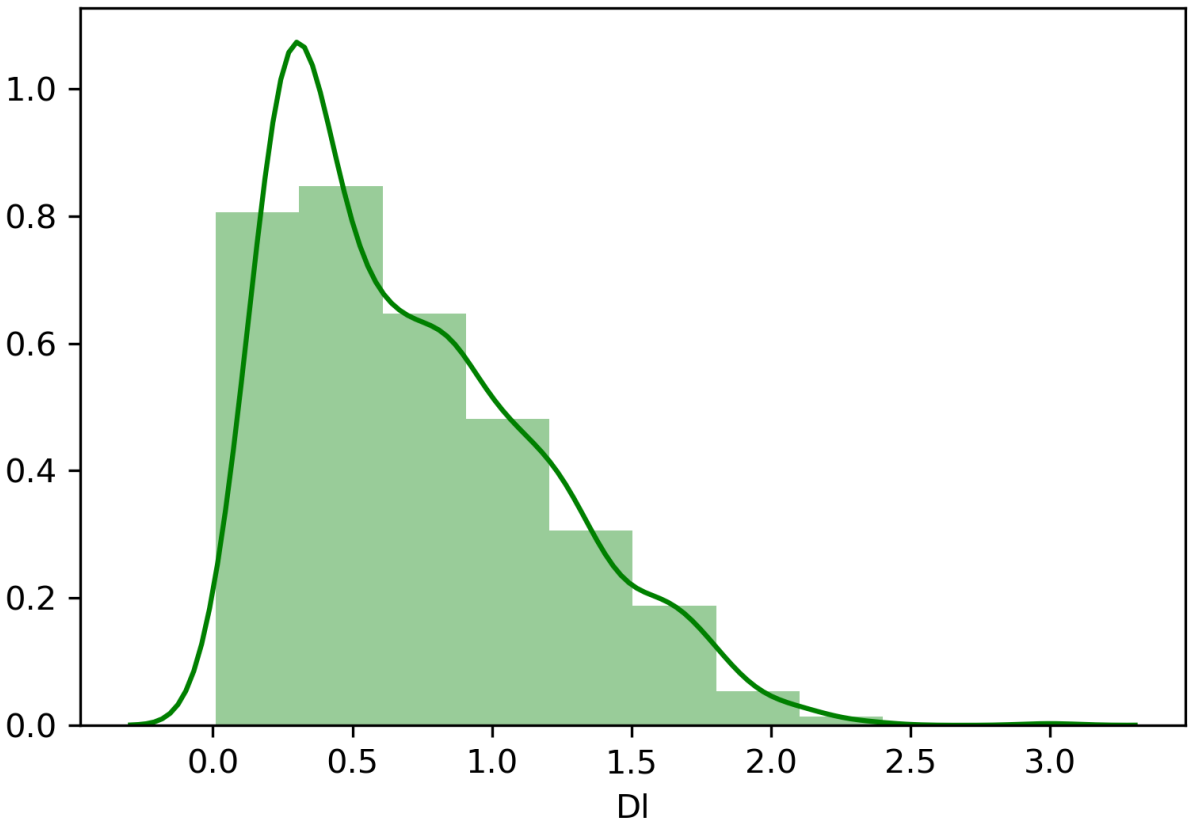
Employment in agriculture (% of total employment)



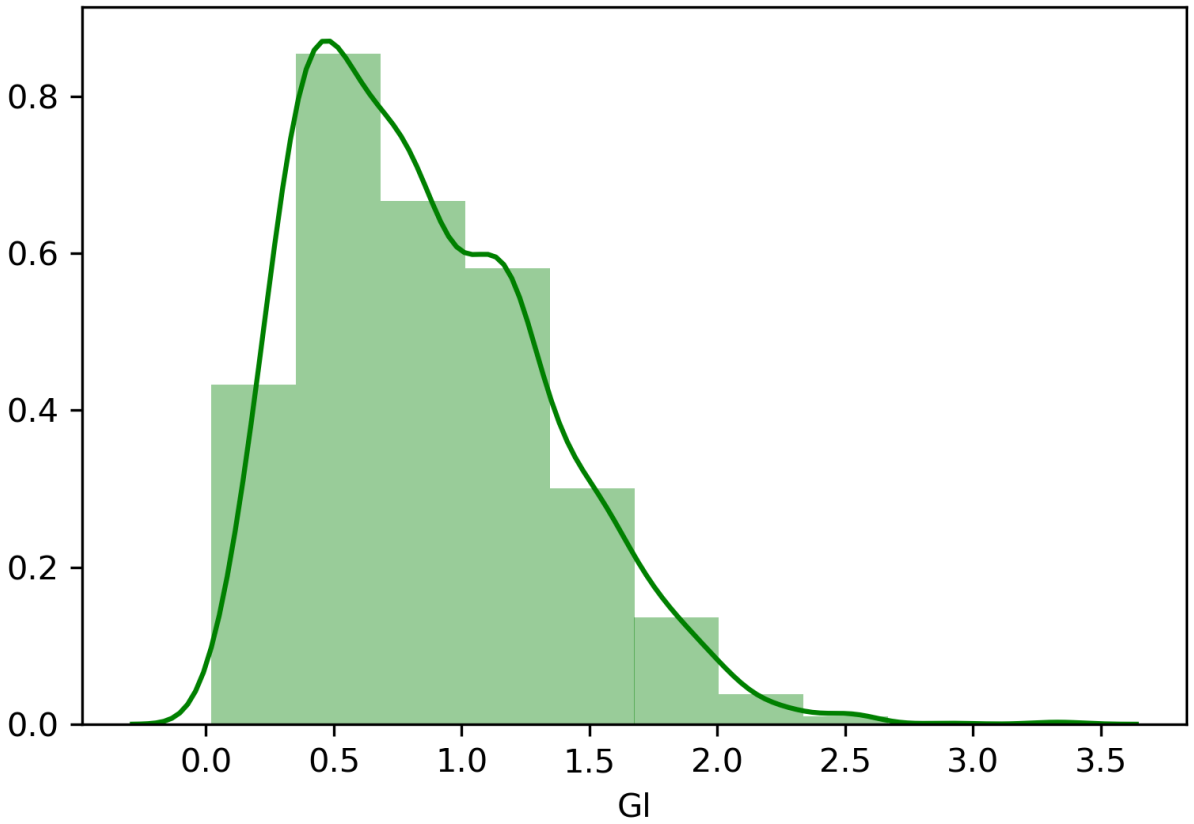
Rural population (% of total population)



Pump price for diesel fuel (US$ per liter)



Pump price for Gasoline fuel (US$ per liter)



GDP per capita (current US$)