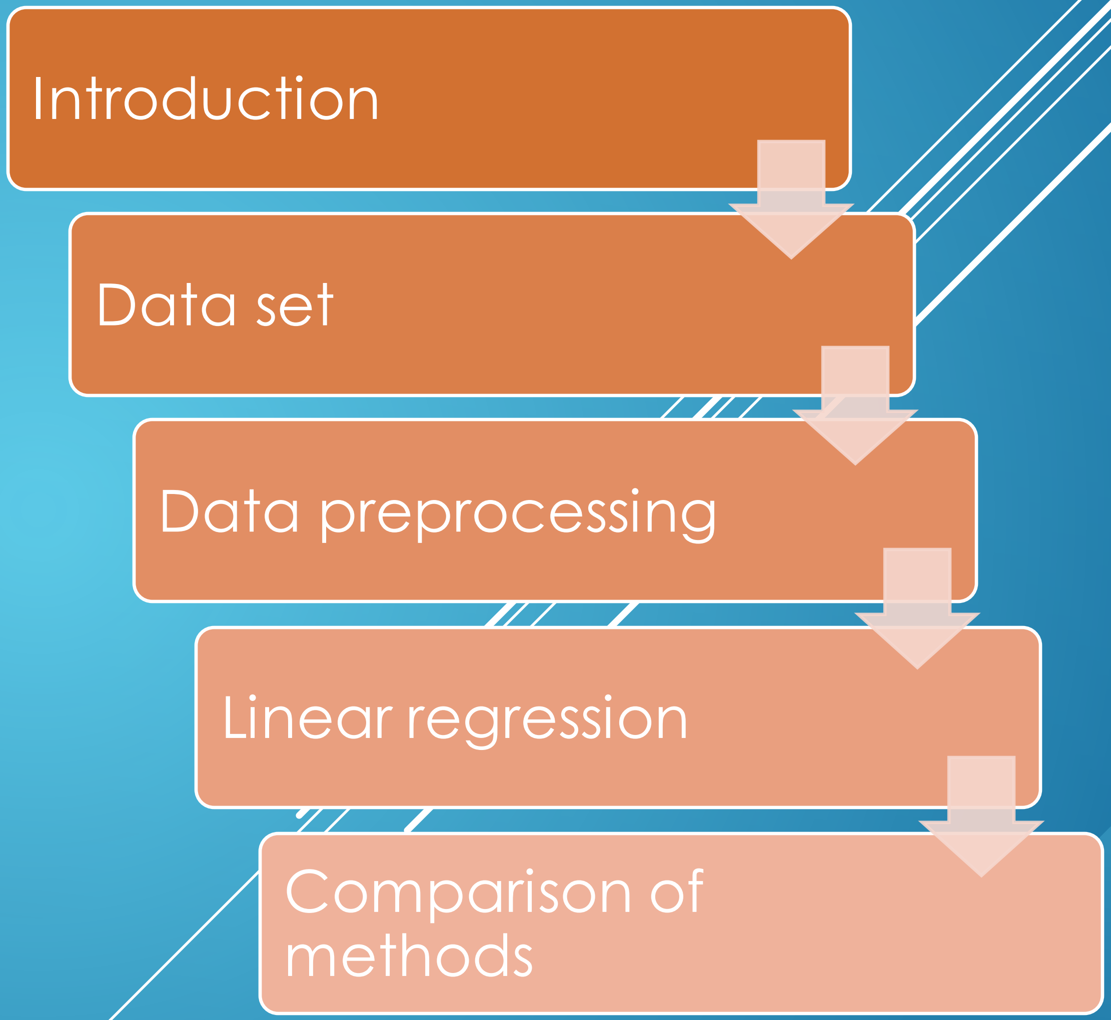
PREDICTING THE NUMBER OF DEATHS CAUSED BY COVID-19

Matyáš Mattanelli

OUTLINE



Introduction

Data set

Data preprocessing

Linear regression

Comparison of methods

INTRODUCTION

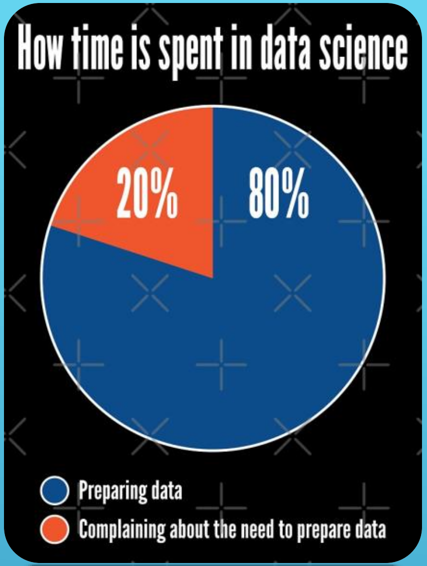
- COVID-19 = Coronavirus Disease 2019
- Caused by coronavirus SARS-CoV-2
  - Severe acute respiratory syndrome coronavirus 2
- First known case in Wuhan (China) in December 2019
- Common symptoms: fever, fatigue, cough, breathing difficulties, loss of smell, and loss of taste
  - 1/3 of infected people do not develop any symptoms at all
  - 3.3% develop critical symptoms (respiratory failure, organ dysfunction)
- Elderly people have a higher risk of developing severe symptoms
- Estimated death rate: 0.99%

DATA SET

Our World in Data

- Obtained from Our World in Data (OWID, https://ourworldindata.org/coronavirus)
- Daily cross-sectional data
- Period: January 2020 – April 2024
- 240 unique locations (“countries”)
- 376 861 observations
- Features
  - Covid-19-related: Number of new deaths, new cases, new vaccinations
  - Country-related: Population, population density, median age, life expectancy, etc.

DATA PREPROCESSING

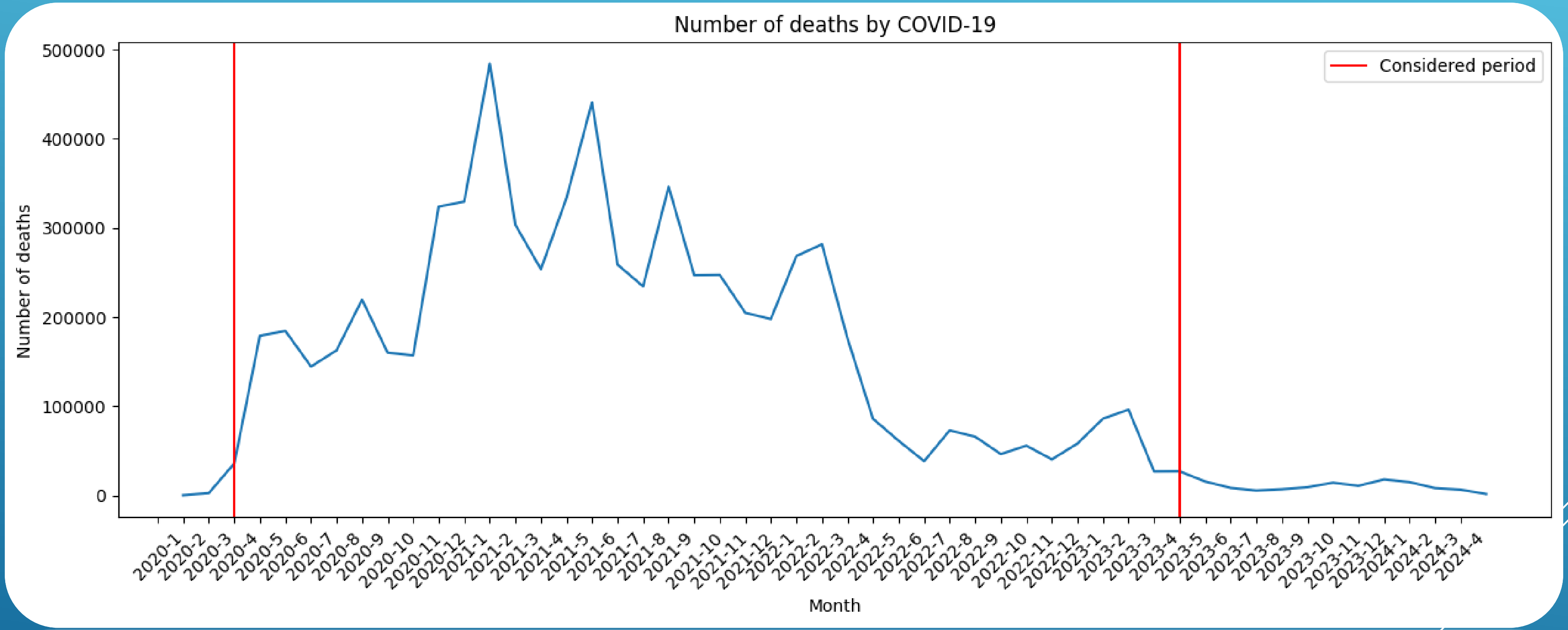


- Daily observations but most of the indicators do not change daily (e.g. population)

  ➡ Data aggregated monthly

- The time span is too wide, the goal is to capture the deaths which happened during the severe stages of the pandemic
  - Time period reduced to March 2020 – April 2023
- Some features have very high percentage of missing values
  - Threshold for preservation set to a maximum of 20% of missing values
- 2 features dropped due to very high correlation with other variables (over 0.9)
- All rows with missing values (after feature filtering) are disregarded
- Final data set with 6 372 observations and 11 features (+ location, date)

NUMBER OF DEATHS OVER TIME (MONTHLY)



DATA PREPROCESSING

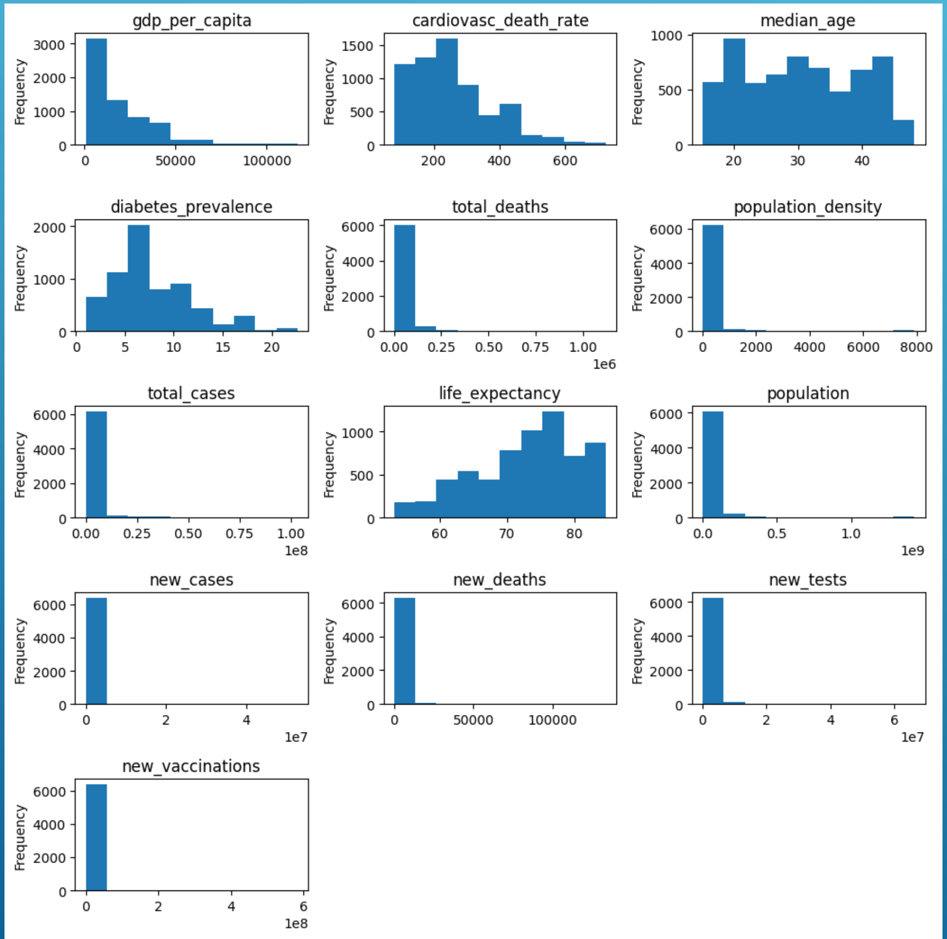
- Most of the features appear to be right-skewed
- If the distribution of the features is non-normal, linear regression may not fit the data well due to its sensitivity to outliers
- Logarithmic transformation may help to make the distribution closer to normal and decrease the effect of outliers
  - Relevant features are transformed in the following way

$$log\_feature = \log(feature + 1)$$

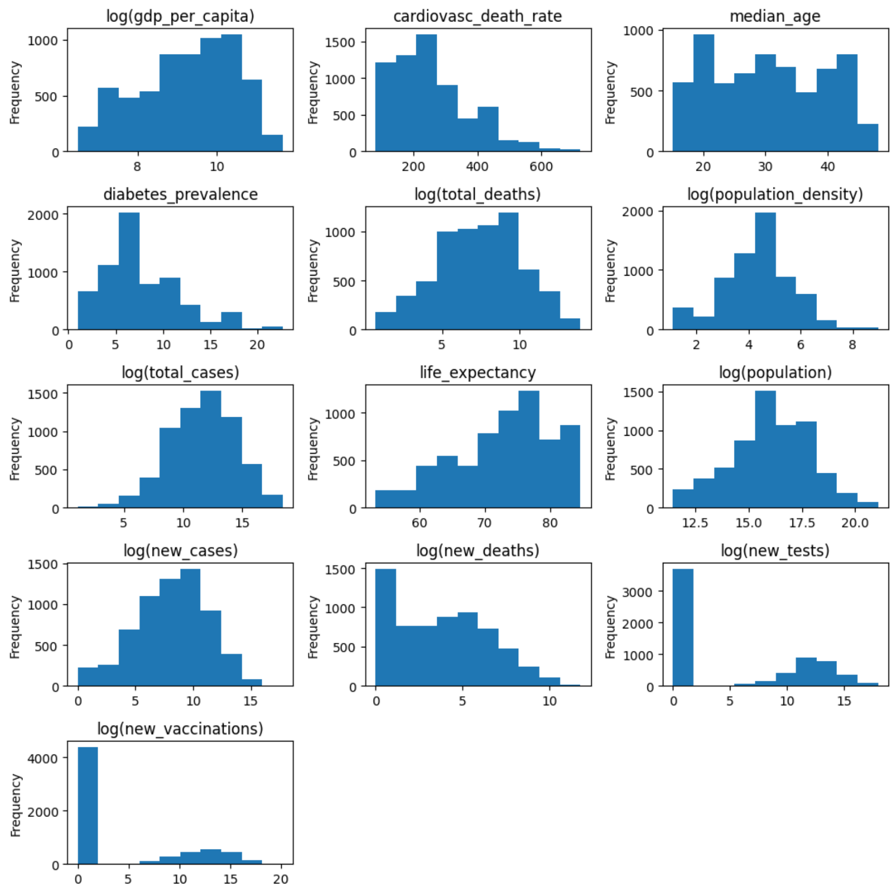
  - 1 is added for features that contain 0

DATA PREPROCESSING

Histograms



Histograms (log)



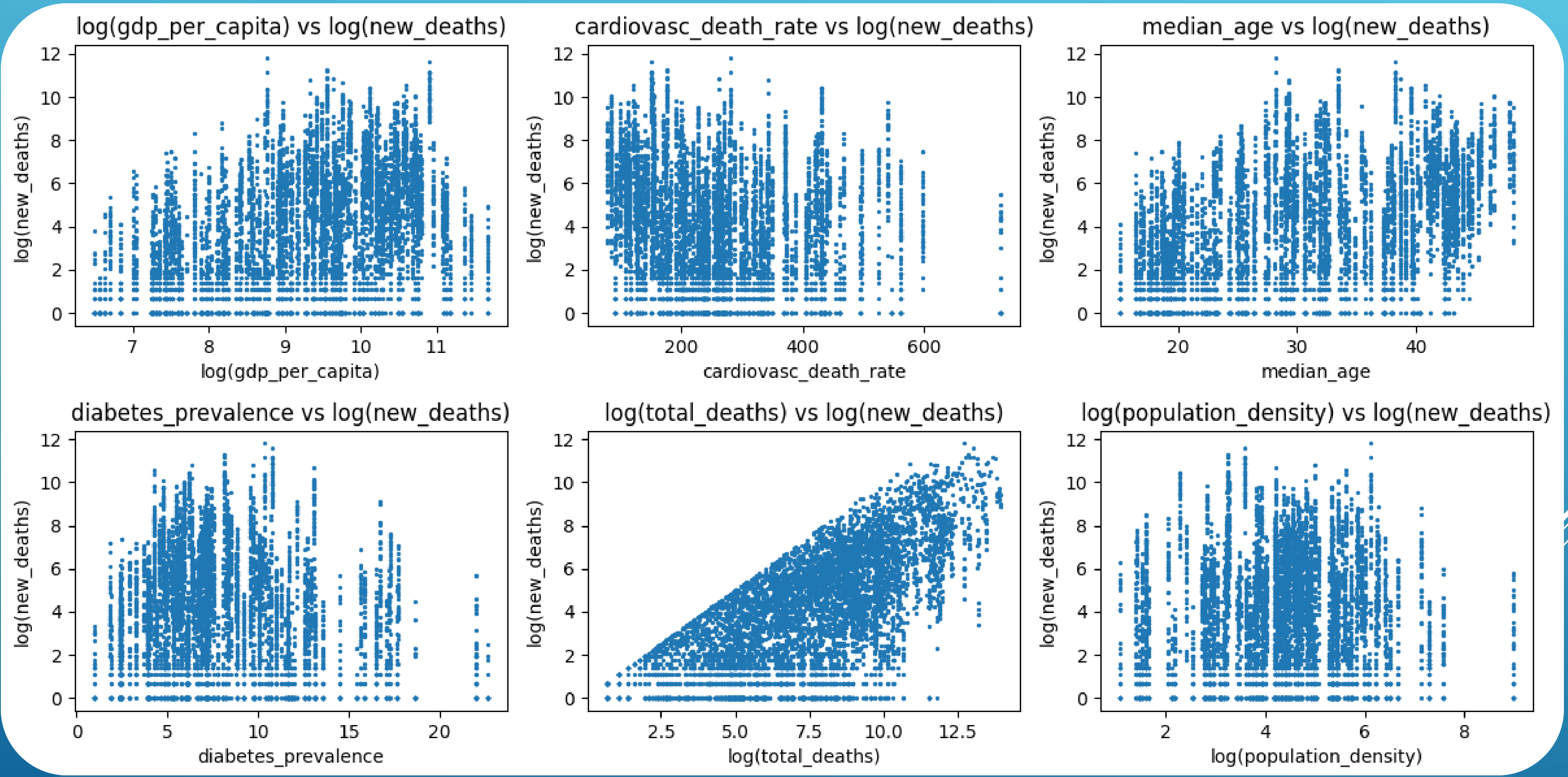
CORRELATION ANALYSIS (INDEPENDENT FEATURES)

Variable 1	Variable 2	Correlation
Log_total_deaths	Log_total_cases	0.93
Median_age	Life_expectancy	0.85
Log_gdp_per_capita	Life_expectancy	0.83
Log_gdp_per_capita	Median_age	0.82
Log_total_cases	Log_new_cases	0.67
Log_total_deaths	Log_new_cases	0.64
Log_total_deaths	Log_population	0.55
Log_new_cases	Log_new_vaccinations	0.49

CORRELATION ANALYSIS (DEPENDENT VARIABLE)

Variable 1	Variable 2	Correlation
Log_new_deaths	Log_new_cases	0.87
	Log_total_deaths	0.67
	Log_total_cases	0.59
	Log_new_tests	0.5
	Log_new_vaccinations	0.47
	Log_population	0.47
	Median_age	0.44
	Life_expectancy	0.41

BIVARIATE ANALYSIS



log(gdp_per_capita) vs log(new_deaths)

cardiovasc_death_rate vs log(new_deaths)

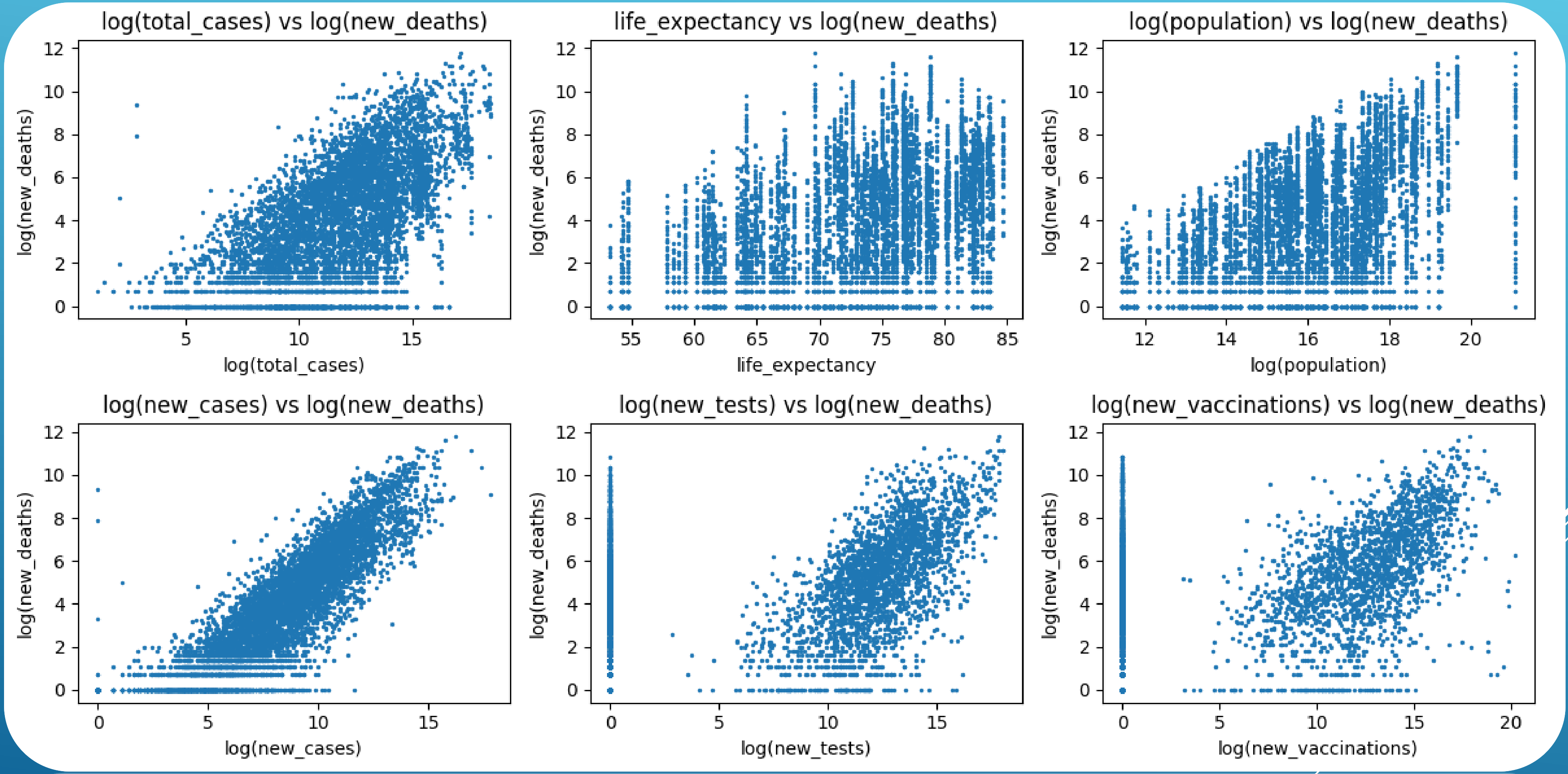
median_age vs log(new_deaths)

diabetes_prevalence vs log(new_deaths)

log(total_deaths) vs log(new_deaths)

log(population_density) vs log(new_deaths)

BIVARIATE ANALYSIS

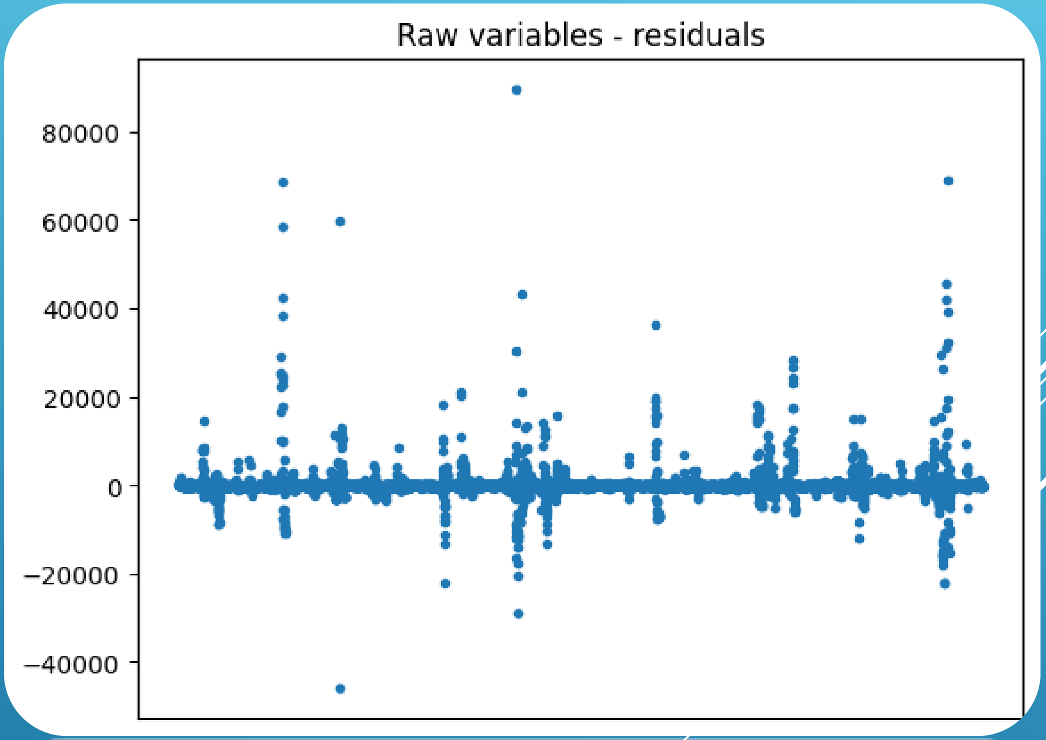


LINEAR REGRESSION – RAW VARIABLES

Variable	Coefficient	Standard error	t-statistic	p-value
Constant	32.26	792.17	0.04	0.97
Gdp_per_capita	-0.007	0.003	-1.87	0.06
Cardiovasc_death_rate	-0.54	0.49	-1.11	0.27
Median_age	26.99	10.18	2.65	0.008
Diabetes_prevalence	6.06	12.73	0.48	0.63
Total_deaths	0.02	0.001	24.16	0.00
Population_density	-0.11	0.07	-1.53	0.13
Total_cases	-0.0002	0.0	-12.94	0.00
Life_expectancy	-5.63	13.52	-0.42	0.13
Population	0.0	0.0	4.73	0.00
New_cases	0.001	0.0	21.24	0.00
New_tests	0.0005	0.0	34.38	0.00
New_vaccinations	1.37e-5	0.0	-3.89	0.00

LINEAR REGRESSION – RAW VARIABLES

Adjusted R-squared	0.438
F-test (statistic)	417.3
F-test (p-value)	0.0
RMSE	3 622.11

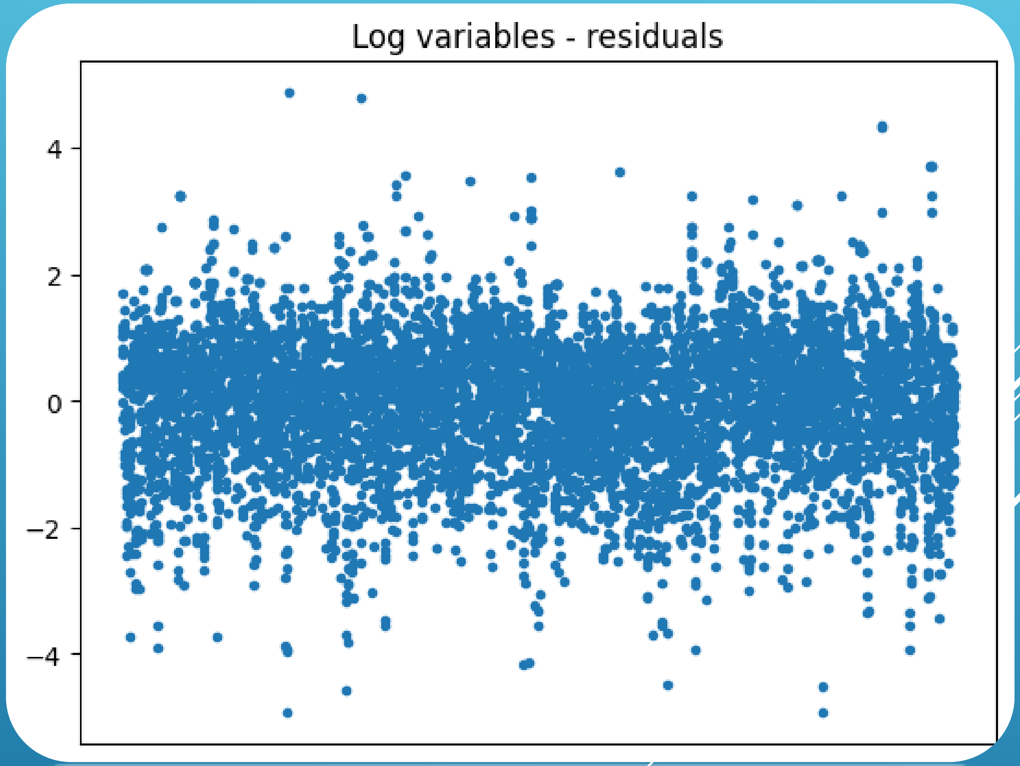


LINEAR REGRESSION – LOG VARIABLES

Variable	Coefficient	Standard error	t-statistic	p-value
Constant	-0.6	0.32	-1.87	0.06
Log_gdp_per_capita	-0.11	0.03	-4.23	0.0
Cardiovasc_death_rate	0.0005	0.0	3.56	0.0
Median_age	0.02	0.003	5.45	0.0
Diabetes_prevalence	0.02	0.004	4.06	0.0
Log_total_deaths	0.76	0.02	49.23	0.0
Log_population_density	-0.05	0.01	-4.12	0.0
Log_total_cases	-0.67	0.02	-44.52	0.0
Life_expectancy	0.01	0.004	2.42	0.02
Log_population	0.04	0.01	4.24	0.0
Log_new_cases	0.65	0.01	93.54	0.0
Log_new_tests	0.04	0.003	13.41	0.0
Log_new_vaccinations	0.02	0.003	5.73	0.0

LINEAR REGRESSION – LOG VARIABLES

Adjusted R-squared	0.846
F-test (statistic)	2943
F-test (p-value)	0.0
RMSE	3561.01

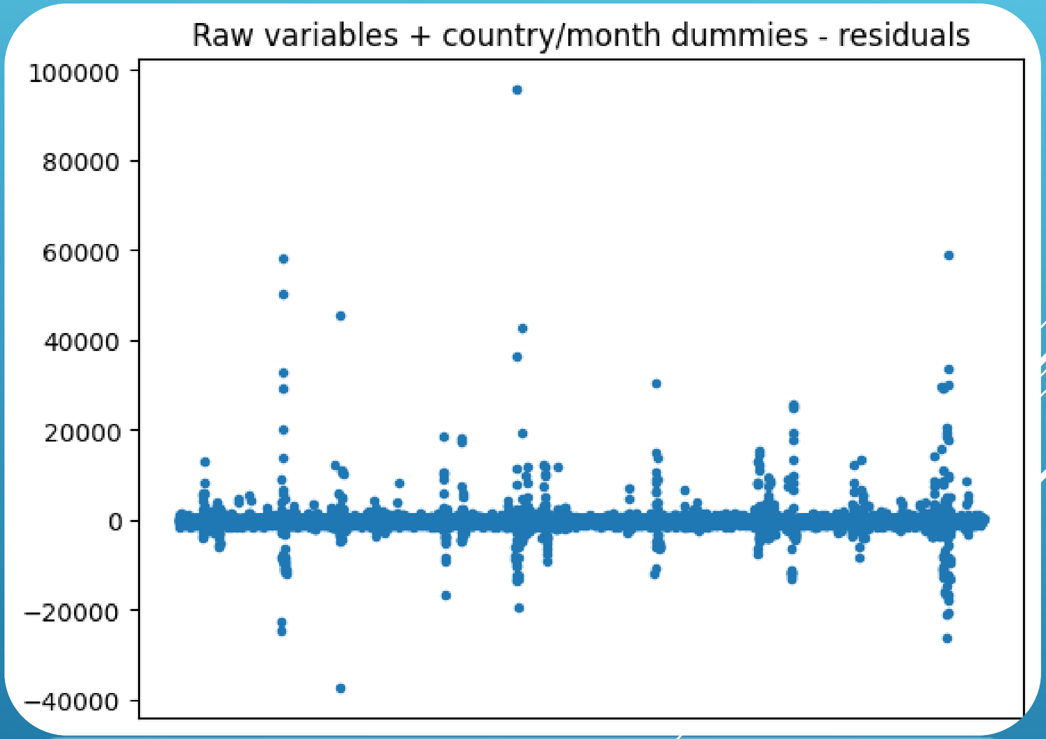


LINEAR REGRESSION – RAW VARIABLES + COUNTRY & DATE DUMMIES

Variable	Coefficient	Standard error	t-statistic	p-value
Constant	-560.92	460.27	-1.22	0.22
Gdp_per_capita	-0.0003	0.005	-0.06	0.95
Cardiovasc_death_rate	-1.05	1.34	-0.78	0.43
Median_age	46.3	12.65	3.66	0.00
Diabetes_prevalence	23.45	11.63	2..02	0.04
Total_deaths	-0.02	0.002	-14.96	0.00
Population_density	-0.47	0.07	-7.05	0.00
Total_cases	6.42e-5	1.42e-5	4.51	0.00
Life_expectancy	-4.64	7.61	-0.61	0.54
Population	7e-6	3.89e-5	18.02	0.00
New_cases	0.0008	5.1e-5	15.9	0.00
New_tests	0.0004	1.65e-5	26.64	0.00
New_vaccinations	2.12e-6	3.26e-6	0.66	0.51

LINEAR REGRESSION – RAW VARIABLES + COUNTRY & DATE DUMMIES

Adjusted R-squared	0.564
F-test (statistic)	38.67
F-test (p-value)	0.0
RMSE	3 138.42

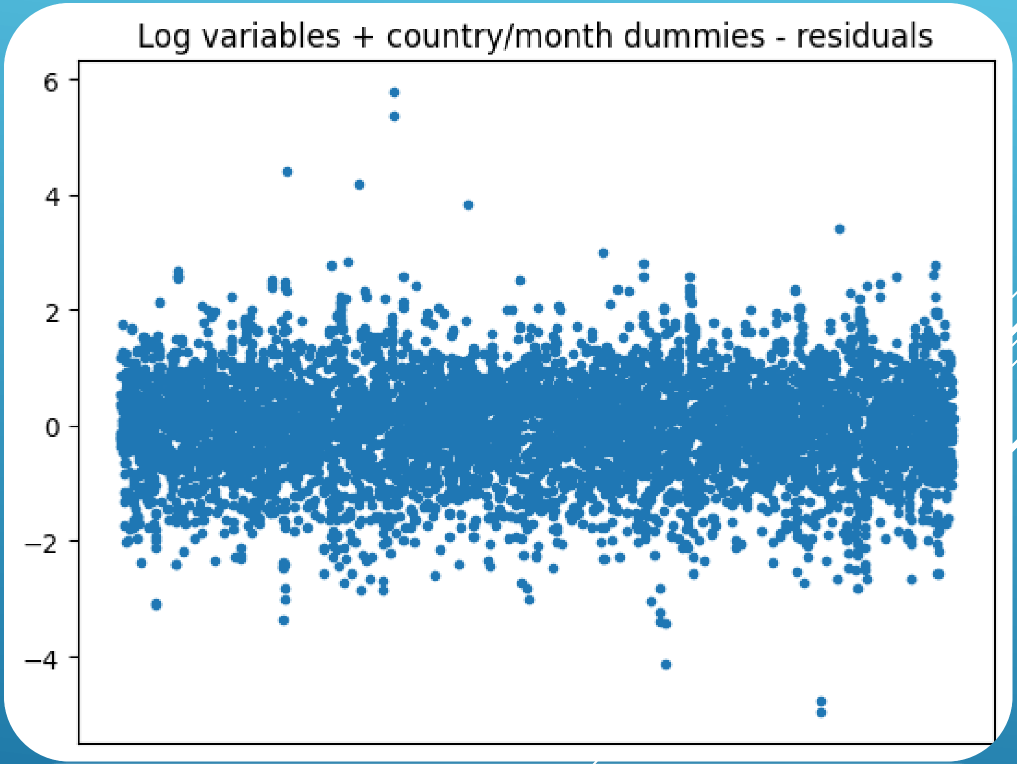


LINEAR REGRESSION – LOG VARIABLES + COUNTRY & DATE DUMMIES

Variable	Coefficient	Standard error	t-statistic	p-value
Constant	1.34	0.18	7.29	0.0
Log_gdp_per_capita	-0.27	0.02	-11.65	0.0
Cardiovasc_death_rate	-1.7e-5	0.0	-0.05	0.96
Median_age	0.002	0.003	0.53	0.6
Diabetes_prevalence	0.02	0.004	6.28	0.0
Log_total_deaths	0.61	0.03	24.28	0.0
Log_population_density	-0.06	0.01	-5.91	0.0
Log_total_cases	-0.1	0.03	-3.6	0.0
Life_expectancy	0.01	0.003	4.56	0.0
Log_population	-0.19	0.02	-12.43	0.0
Log_new_cases	0.53	0.009	62.03	0.0
Log_new_tests	0.01	0.003	3.44	0.001
Log_new_vaccinations	0.02	0.003	7.87	0.0

LINEAR REGRESSION – LOG VARIABLES + COUNTRY & DATE DUMMIES

Adjusted R-squared	0.894
F-test (statistic)	245.5
F-test (p-value)	0.0
RMSE	2195.68



LINEAR REGRESSION - SUMMARY

	Adjusted R-squared	RMSE
Raw variables	0.438	3 622.11
Log variables	0.846	3 561.01
Raw variables + dummies	0.564	3 138.42
Log variables + dummies	0.894	2 195.68

PREDICTIVE ANALYSIS

- So far only fit on the training data was considered
  - 80/20 split to evaluate performance on the test set
- Comparison of multiple methods
  - Linear regression (No regularization, Ridge, Lasso)
  - SVM
  - Random Forest
  - MLP
- Grid search (for relevant methods) through 3-fold cross-validation
- Evaluation metric: RMSE

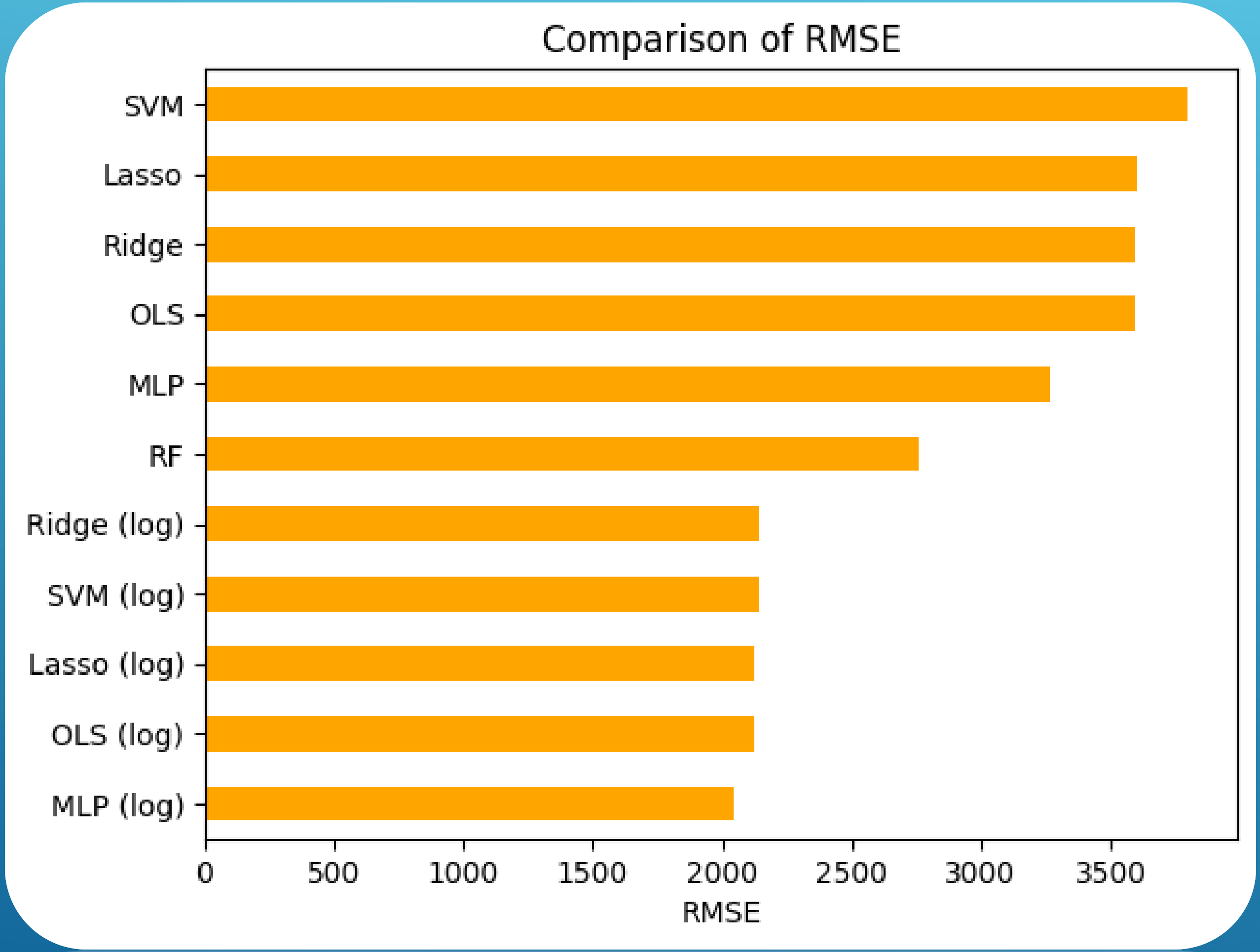
PREDICTIVE ANALYSIS – BEST PARAMETERS

- Ridge
    - Regularization strength = 10
- Lasso
    - Regularization strength = 5
- SVM
    - Regularization strength = 1/50
    - Kernel = polynomial
    - Degree = 3
    - Coef0 = 5
    - Gamma = 1/no_of_features

PREDICTIVE ANALYSIS – BEST PARAMETERS

- Random Forest
  - Max depth = 30
  - Fraction of features considered during split: 0.5
  - Number of trees = 500
- MLP
  - Regularization strength = 0.001
  - Number of nodes in the hidden layer = 100

PREDICTIVE ANALYSIS - RESULTS



THANK YOU FOR YOUR ATTENTION