Documentation for Julia Machine Learning Code

Packages Used:

DataFrames: Provides tools for working with structured data.

CSV: Utilized for reading and writing CSV files.

GLM (Generalized Linear Models): Used for fitting generalized linear models.

Plots and StatsPlots: Used for creating visualizations.

LinearAlgebra: Provides a suite of functions for matrix operations.

Distributions: Used for working with statistical distributions.

Key Code Analysis and Results:

The notebook involves loading a dataset 'clouds.csv' using CSV.read(), directly into a DataFrame. This dataset includes variables like seeding, cloud cover, and rainfall among others.

Visual Analysis:

Boxplots and scatter plots are generated to visualize the relationship between rainfall and other variables such as 'seeding' and 'echomotion'. These plots are helpful in identifying patterns, outliers, and distribution shapes.

Regression Model:

A multiple linear regression model is fitted using a specific formula with interaction terms. The output shows coefficients for each predictor, indicating their impact on rainfall.

Statistical Measures:

Residual Analysis: Residual plots are used to assess the model's fit, including a Q-Q plot to check the normality of residuals.

Cook's Distance: Used for identifying influential cases that might affect the regression model's estimates.

MSE (Mean Squared Error): Calculated to quantify the average squared difference between the observed actual outcomes and the outcomes predicted by the model.

Diagnostic Plots:

Several diagnostic plots like residuals vs. fitted values and leverage points are plotted to check for model assumptions and potential influential observations.

Regression Results:

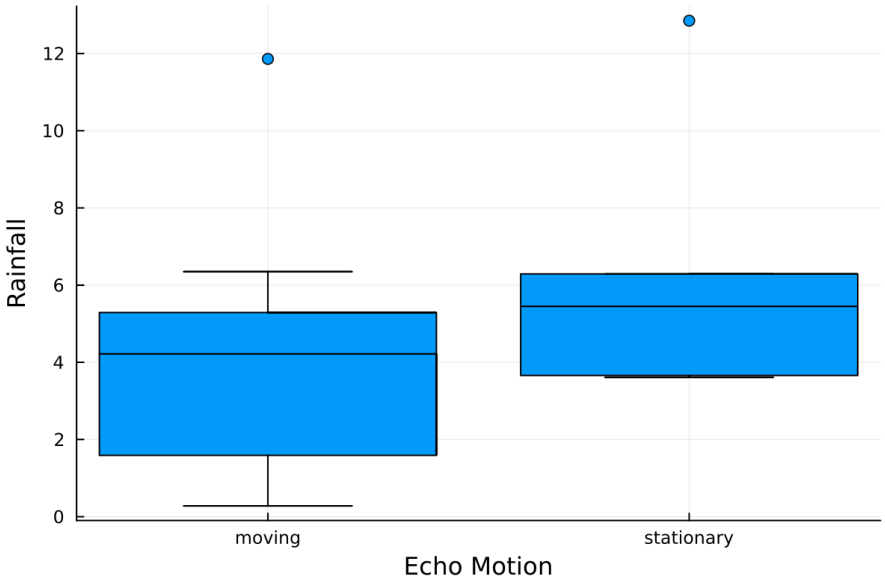
Variable	Coefficient	Std. Error	t-value	Pr(>|t|)	95% Confidence Interval
(Intercept)	-0.346	2.788	-0.12	0.903	[-6.369, 5.676]
seeding: yes	15.683	4.446	3.53	0.004	[6.077, 25.289]
sne	0.420	0.845	0.50	0.627	[-1.405, 2.244]
cloudcover	0.388	0.218	1.78	0.098	[-0.083, 0.859]
prewetness	4.108	3.601	1.14	0.275	[-3.671, 11.888]
echomotion: stationary	3.153	1.933	1.63	0.127	[-1.022, 7.328]
time	-0.045	0.025	-1.80	0.096	[-0.099, 0.009]
seeding: yes & sne	-3.197	1.267	-2.52	0.025	[-5.935, -0.460]
seeding: yes & cloudcover	-0.486	0.241	-2.02	0.065	[-1.007, 0.035]
seeding: yes & prewetness	-2.557	4.481	-0.57	0.578	[-12.238, 7.123]
seeding: yes & echomotion: stationary	-0.562	2.644	-0.21	0.835	[-6.275, 5.150]

Visualizations

The following figures display various visualizations generated from the Cloud Seeding data analysis. These include diagnostic plots, scatterplots, and boxplots to better understand the relationships and residuals.

1. 3. Boxplot

This figure presents a boxplot of rainfall by seeding category (No seeding vs Seeding).



2. Scatterplots

Scatterplot of Rainfall Against Time

The scatterplot plots rainfall against time to visualize trends or patterns over the period studied.

Scatterplot of Rainfall Against Cloud Cover

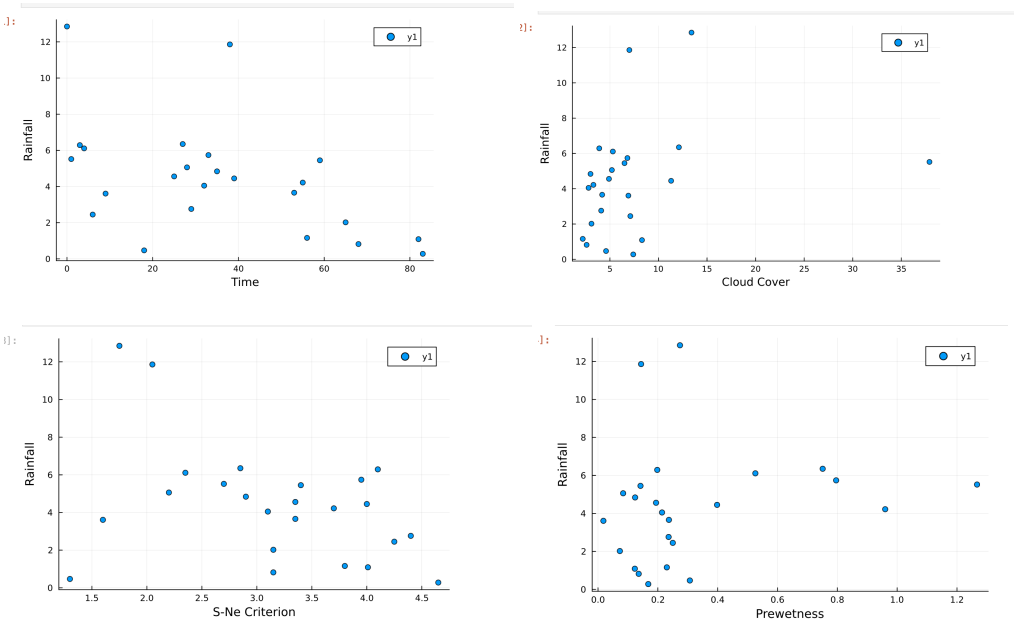
This scatterplot shows how rainfall varies with cloud cover.

Scatterplot of Rainfall Against S-Ne Criterion

Rainfall is plotted against the S-Ne criterion to investigate any correlation between this atmospheric measurement and rainfall levels.

Scatterplot of Rainfall Against Prewetness

This plot visualizes the relationship between the prewetness of the atmosphere and the rainfall.



3. Diagnostic Plots

Residuals vs. Fitted Values Plot

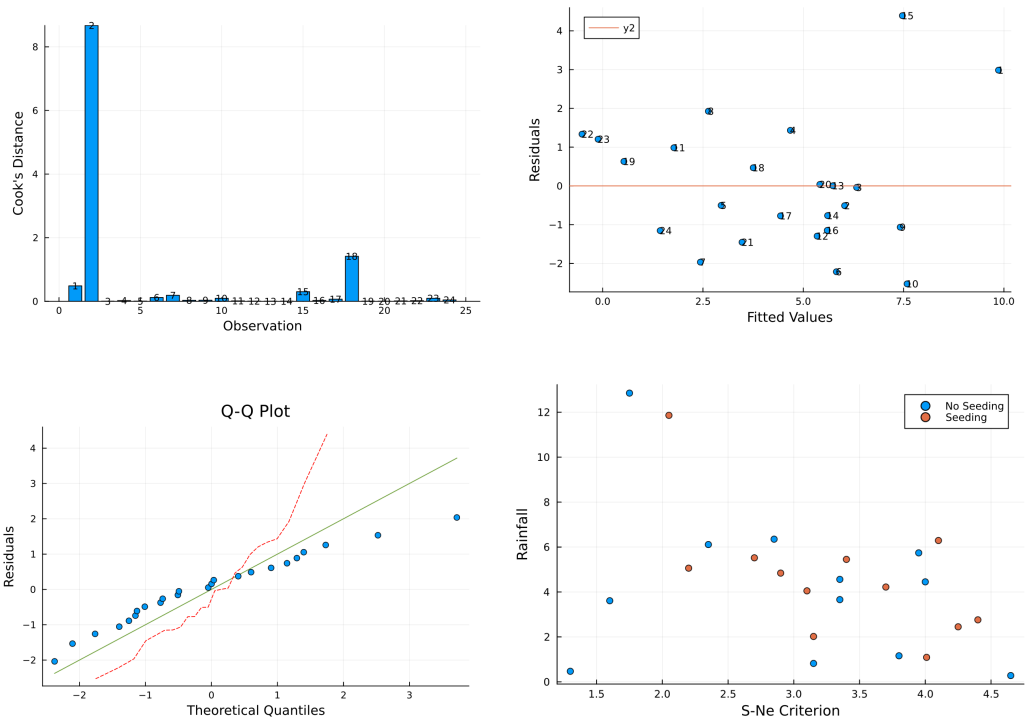
Used to assess the goodness of fit for the regression model.

Cook's Distance Plot

This plot identifies influential data points that might have a significant impact on the calculation of regression coefficients.

Q-Q Plot of Residuals

Used to check the normality assumption of the residuals.



Mac OS (Jupyter notebooks: 1.3 sec , Pluto jl:0.9 sec)