

STA 325 Case Study

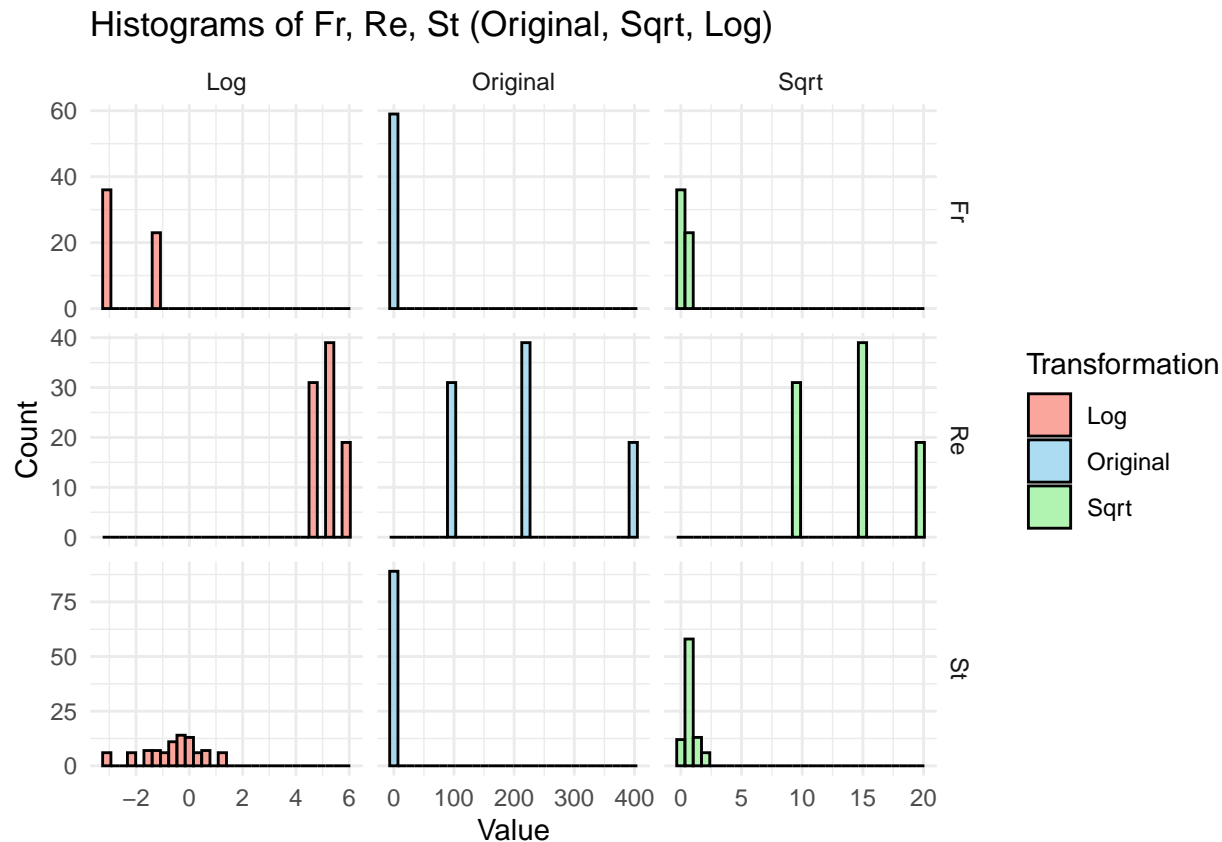
Abby L., Ava E., Ella T., Grady P. Laura C.

Introduction

Whether it's stirring a morning cup of coffee or feeling your stomach lurch during a bumpy airplane ride, nearly everyone has experienced some form of turbulence. Beyond these everyday encounters, turbulence plays a crucial role in many complex natural and industrial processes, from air pollution and chemical reactions to heat transfer and weather systems. Despite its ubiquity, turbulence remains notoriously difficult to predict and has long been regarded as "the last great unsolved problem in classical physics."

While we may not be physicists ourselves, as members of Stat 325, we have undertaken the challenge of developing a predictive model to better understand this elusive phenomenon in collaboration with our Professor, Simon Mak. Given observations for fluid turbulence (quantified by Reynolds number Re), gravitational acceleration (quantified by Froude number Fr), and the particle's characteristics (quantified by Stokes number St), we have explored models designed to predict the mean, standard deviation, skewness, and kurtosis of the spatial distribution and clustering of particles in clouds in a state of idealized turbulence. Our goal is to create a model that balances predictive accuracy with interpretability; one that not only performs well statistically but also provides clear, meaningful insights into how different conditions influence turbulent behavior.

Exploratory Data Analysis



We observed that almost all of the residual models have clustered residuals, likely due to both Fr and Re only having 3 unique levels in the data.

Methodology

Observing the distributions of the response variables, the square root of St was used as well as the log of Re to improve the symmetry of the distribution to address the constant variance assumption. Additionally, the logit function was applied to Fr so that all values are bound between 0 and 1.

One concern with the data is that Re and Fr each only have 3 unique values. Hence, using these variables to model the distribution of the response variable requires a large amount of uncertainty and interpolation at values of these variables between the 3 distinct values in the data. To continue refining the models, additional data with different values for these models will reduce uncertainty and improve model fit.

Mean

To predict the mean particle cluster volume, we assessed the performance of a few different models including simple linear regression, interaction terms, polynomial models, ridge regression, and spline. We then assessed the fit of each by performing 5-fold CV. The polynomial model performed best. Then, we performed a log transform on the response variable. Our final model is the polynomial regression model on the log-transformed mean. The log transformation was necessary because without it, the residuals showed clear patterns and heteroscedasticity. After transformation, the residuals appear to be randomly scattered around

zero with constant variance. The log transformation also addressed the right-skewed distribution of mean cluster volumes in the raw data.

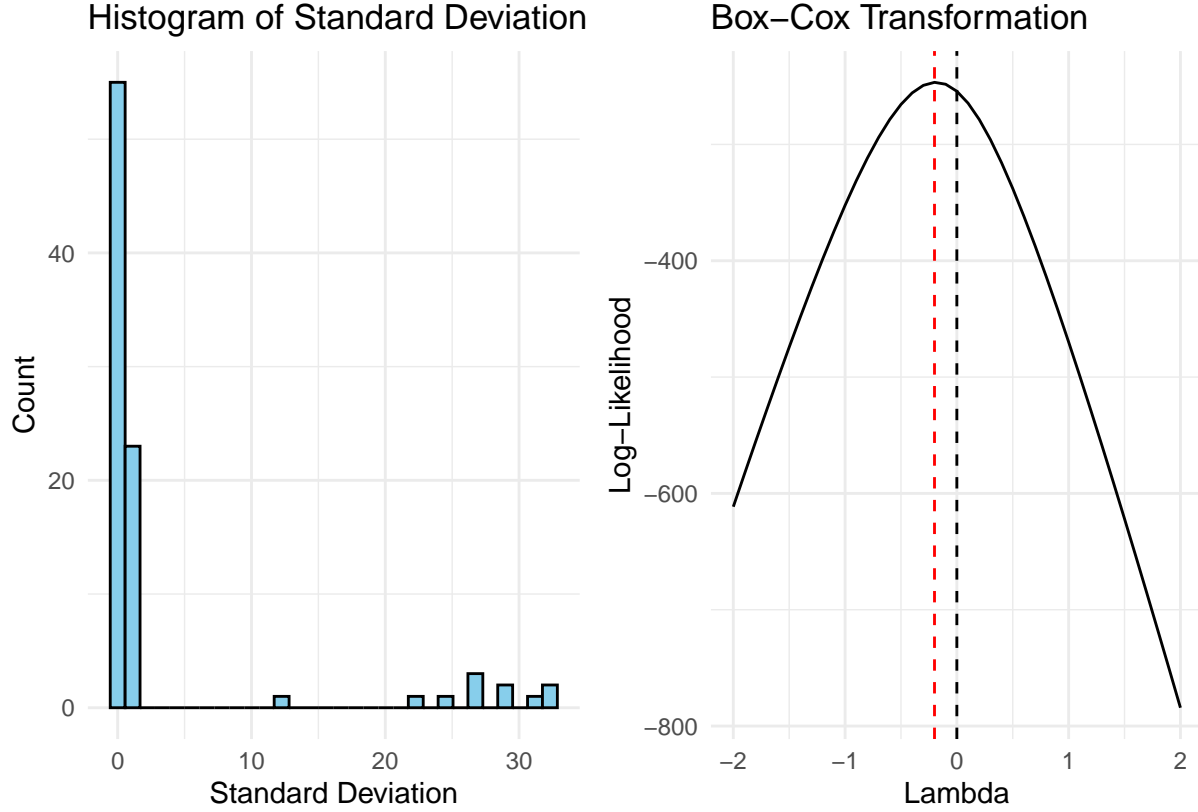
$$\log(\text{mean}) = \beta_0 + \beta_1 \text{poly}_1(\sqrt{St}) + \beta_2 \text{poly}_2(\sqrt{St}) + \beta_3 \text{poly}_1(Re) + \beta_4 \text{poly}_2(Re) + \beta_5 \text{poly}_1(\text{logisticFr}) + \beta_6 \text{poly}_2(\text{logisticFr}) + \beta_7 (\text{St:Re}) + \beta_8 (\text{St:logisticFr}) + \beta_9 (\text{Re:logisticFr}) + \beta_{10} (\text{St:Re:logisticFr})$$

The final model includes second-degree polynomial terms for all three predictors (\sqrt{St} , Re , and the logistic-transformed Fr) along with three two-way interactions: $\sqrt{St}:Re$, $\sqrt{St}:\text{logistic_Fr}$, and $Re:\text{logistic_Fr}$. This model achieved great fit, with an adjusted R^2 of 0.9978, meaning it explains over 99% of the variation in $\log(\text{mean})$ cluster volume. The model's RMSE of 0.098 on the log scale translates to predictions that are typically within about 10% of the true mean cluster volume when back-transformed, although there are still some uncertainty regarding the predictions due to the presence of categorical predictors.

We also ran ANOVA which reveals which predictors contribute most substantially to explaining mean cluster volume. The results show that Reynolds number has by far the largest effect (Sum of Squares = 431.68), followed by Stokes number (SS = 7.53), and then Froude number (SS = 0.19). All three interaction terms were statistically significant ($p < 0.05$), with the $Re:\text{logistic_Fr}$ interaction contributing the most (SS = 0.48, $p < 0.001$). The significant $\sqrt{St}:\text{logistic_Fr}$ interaction ($p = 0.001$) and marginally significant $\sqrt{St}:Re$ interaction ($p = 0.030$) suggest that the effects of particle inertia (St) on mean cluster size depend on both gravitational and turbulence conditions. To further assess whether this complex model overfits the data, we performed 5-fold cross-validation, averaging the RMSE across folds to estimate out-of-sample prediction error. The cross-validated RMSE was 0.118, only slightly higher than the training RMSE of 0.098, and the CV R^2 (0.9974) remained nearly identical to the training R^2 (0.9978). This close agreement between training and validation performance indicates that the model generalizes well to unseen data and does not appear to overfit the data despite its polynomial and interaction terms.

Overall, from fitting the mean with a polynomial model, we see that Reynolds number shows a strong negative linear effect and positive quadratic effect, indicating that mean cluster volume initially decreases sharply with increasing turbulence intensity, then levels off at higher Re values. The Stokes number exhibits a similar but weaker pattern—cluster volumes first increase then decrease with particle inertia. The Froude number's negative linear and positive quadratic effects suggest that gravitational acceleration initially reduces clustering but this effect diminishes at extreme gravity levels. The significant $Re:Fr$ interaction indicates that gravity's effect on clustering becomes stronger in more turbulent flows, which makes physical sense as turbulence and gravity compete in determining particle settling behavior. Overall, we found that Reynolds number (turbulence intensity) is by far the dominant factor controlling mean cluster size, but its effect depends significantly on gravitational conditions, specifically gravity's impact on clustering strengthens in more turbulent flows. This interaction means that predicting particle behavior requires considering both turbulence and gravity together, rather than treating them as independent effects.

Standard Deviation



The Box Cox transformation shows that the optimal lambda is around 0, so the log transformation on sd is the optimal transformation to address skewness of the response variable.

We were given a training and testing dataset, so I trained different linear (with and without ridge and lasso), polynomial and spline regression models comparing the adjusted R-squared the p-values, and used 5-fold CV to compare the top model to determine which model had the best prediction accuracy while working on new, unseen data.

After multiple models were compared from their adjusted R-squared and 5-fold CV, the linear model with all of the interaction terms was selected based on the balance between predictive accuracy (RMSE = 2.01, Adjusted R-squared = 0.74). The interaction terms were included to capture the potential effects of predictors coupled together. The linear model is interpretable and less risky to overfit, which is helpful in our goal to predict the standard deviations of the distributions given St, Re, and Fr.

$$Fr_{logistic} = \frac{1}{1 + \exp(-Fr)} \log(\hat{sd}) = 26.5 + 0.642 \log(St) - 4.915 \log(Re) - 20.704 Fr_{logistic} - 0.027 \log(St) \log(Re) - 0.102 \log(St) Fr_{logistic}$$

Model	RMSE
lin	2.077942
lin_interactions	2.013816
lin_interactions_subset	2.166138
poly	2.072878
spline	2.115911

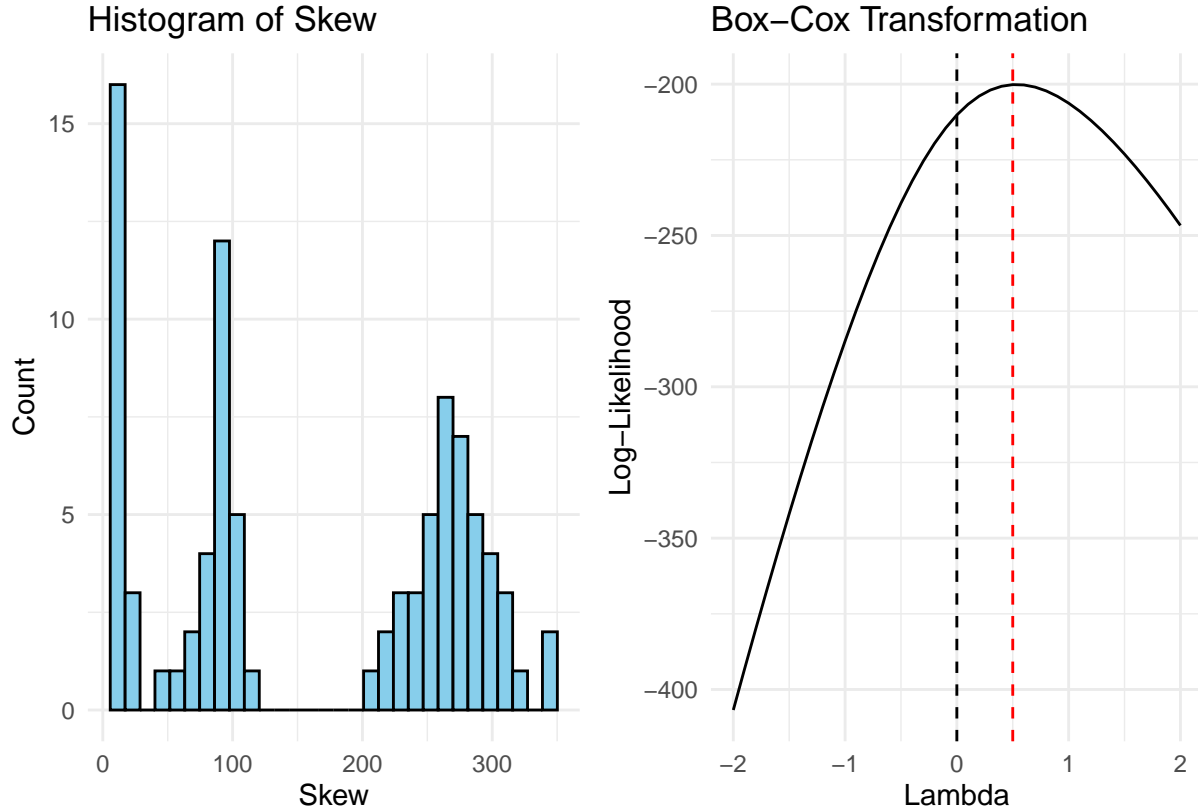
Model	RMSE
lasso	2.068582
ridge	2.080013

When running 5-fold CV, the RMSE was 2.0138, which lower compared to other linear, ridge, lasso, spline and polynomial models.

```
##
## Call:
## lm(formula = log(sd) ~ (log_St + log_Re + logistic_Fr)^2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35706 -0.43689  0.09339  0.41670  1.90516
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26.49639     3.06941   8.632 3.88e-13 ***
## log_St         0.64222     0.89622   0.717  0.476
## log_Re        -4.91459     0.58036  -8.468 8.22e-13 ***
## logistic_Fr    -20.70407     4.25749  -4.863 5.50e-06 ***
## log_St:log_Re   -0.02716     0.16702  -0.163  0.871
## log_St:logistic_Fr -0.10234     0.41962  -0.244  0.808
## log_Re:logistic_Fr  3.55720     0.79835   4.456 2.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9463 on 82 degrees of freedom
## Multiple R-squared:  0.7577, Adjusted R-squared:  0.74
## F-statistic: 42.75 on 6 and 82 DF,  p-value: < 2.2e-16
```

Skew

Looking at the distribution of a Skew through the EDA and Box-Cox transformation (below), a square root transformation on Skew was deemed most appropriate to address the right skewness of the response variable.

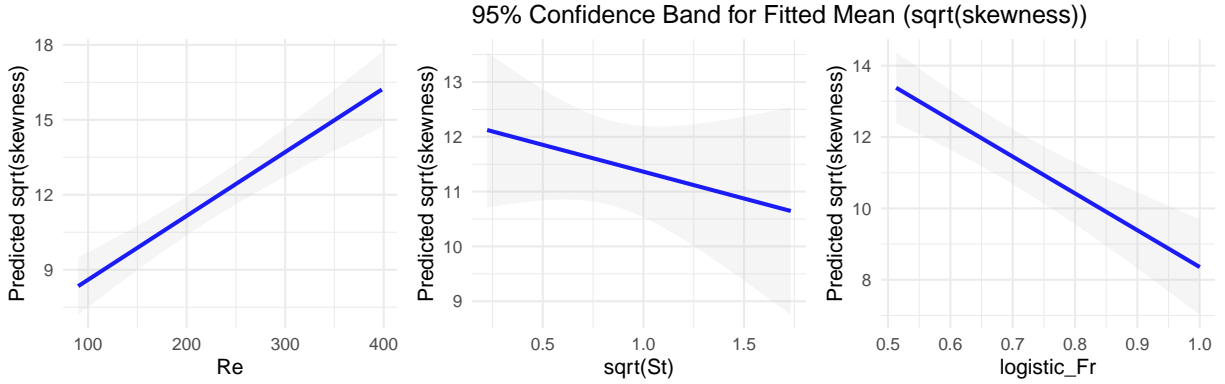


The model which best fit the $\sqrt{\text{skewness}}$ response is a ridge model fit with 5 fold cross validation. To evaluate the significance of coefficients, the same model was then fit with OLS to develop better intuition regarding the magnitude and significance of coefficients. The below table highlights two of the most significant coefficients:

Predictor	Estimate
logistic_Fr	-21.350
Re:logistic_Fr	0.055

NOTE: Need to add interpretation of the coefficients here

Given the nature of the three levels for both Re and Fr, there is a large degree of uncertainty for these predictors when predicting the response which is especially evident in the wide nature of the 95% confidence intervals near the extremes of both of these predictors. The below plots show slices of each of the predictors for the response at the mean of the other two predictors which aren't on the x-axis. Receiving additional data in the future would address the high amount of uncertainty in the model as we continue to refine the model.



Results

Standard Deviation

CONFIDENCE INTERVAL

```
##
## Call:
## lm(formula = log(sd) ~ (log_St + log_Re + logistic_Fr)^2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35706 -0.43689  0.09339  0.41670  1.90516
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26.49639     3.06941   8.632 3.88e-13 ***
## log_St         0.64222     0.89622   0.717  0.476
## log_Re        -4.91459     0.58036  -8.468 8.22e-13 ***
## logistic_Fr    -20.70407     4.25749  -4.863 5.50e-06 ***
## log_St:log_Re  -0.02716     0.16702  -0.163  0.871
## log_St:logistic_Fr -0.10234     0.41962  -0.244  0.808
## log_Re:logistic_Fr  3.55720     0.79835   4.456 2.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9463 on 82 degrees of freedom
## Multiple R-squared:  0.7577, Adjusted R-squared:  0.74
## F-statistic: 42.75 on 6 and 82 DF,  p-value: < 2.2e-16
```

From our model, we can see that \log_Re , \logistic_Fr , and the interaction between \log_Re and \logistic_Fr has significant effects on the standard deviation of the particle cluster volume distribution. For example, holding other factors constant, for a 1% increase in the Reynold's Number, there would be a 4.915% decrease standard deviation. Also, holding other factors constant, for a 1% increase in the Stokes number, there would be a 20.704% decrease in standard deviation. In context, this means that more turbulent regimes and regimes with more dense particles lead to narrower particle cluster distributions. However, the effect with Reynold's Number is reduces in regimes with higher gravitational acceleration, or Fr , as seen with the interaction effect between $\log(Re)$ and $\logistic(Fr)$.

Predictions on Test Data

##	St	Re	Fr	logistic_Fr	sd
## 1	0.05	398	0.052	0.5129971	-3.90090520
## 2	0.20	398	0.052	0.5129971	-3.30881187
## 3	0.70	398	0.052	0.5129971	-2.77375048
## 4	1.00	398	0.052	0.5129971	-2.62141281
## 5	0.10	398	Inf	1.0000000	-3.20229953
## 6	0.60	398	Inf	1.0000000	-2.52633480
## 7	1.00	398	Inf	1.0000000	-2.33361914
## 8	1.50	398	Inf	1.0000000	-2.18065212
## 9	3.00	398	Inf	1.0000000	-1.91915327
## 10	3.00	224	0.300	0.5744425	-0.45526143
## 11	0.10	224	Inf	1.0000000	-2.45801606
## 12	0.50	224	Inf	1.0000000	-1.82570540
## 13	0.40	90	0.052	0.5129971	1.54359584
## 14	1.00	90	0.052	0.5129971	1.97194993
## 15	0.05	90	0.300	0.5744425	0.30169262
## 16	0.30	90	0.300	0.5744425	1.12804942
## 17	0.60	90	0.300	0.5744425	1.44772787
## 18	0.80	90	0.300	0.5744425	1.58040641
## 19	0.40	90	Inf	1.0000000	-0.69835198
## 20	0.50	90	Inf	1.0000000	-0.60515716
## 21	0.60	90	Inf	1.0000000	-0.52901145
## 22	1.50	90	Inf	1.0000000	-0.14632707
## 23	2.00	90	Inf	1.0000000	-0.02617805

Conclusion