

FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION
OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice #2

ANALYSIS OF MULTIVARIATE RANDOM VARIABLES

Performed by
Dmitry Grigorev,
Eugenia Khomenko,
Efim Podkovirkin,
Arina Syrchenko

St. Petersburg

2022

Contents

| | | |
|------|---|-----------|
| 1. | Data description | 3 |
| 2. | Plotting a non-parametric estimation of PDF in form of a histogram and kernel density function for MRV (or probability law in case of discrete MRV) | 4 |
| 3. | Estimation of multivariate mathematical expectation and variance | 7 |
| 4. | Non-parametric estimation of conditional distributions, mathematical expectations and variances | 9 |
| 5. | Estimation of pair correlation coefficients, confidence intervals for them and significance levels | 10 |
| 5.1. | Theory | 10 |
| 5.2. | Results | 11 |
| 6. | Task formulation for regression | 12 |
| 7. | Regression model construction | 12 |
| 7.1. | Model 1 | 12 |
| 7.2. | Multicollinearity and partial correlation analysis | 12 |
| 7.3. | Outlier with respect to distribution removal | 14 |
| 7.4. | Model 2 | 15 |
| 7.5. | Outliers with respect to regression analysis | 17 |
| 7.6. | Model 3 | 18 |
| 8. | Appendix | 19 |
| | Bibliography | 20 |

1. Data description

Let D is a subsample on size $n = 1000$ from modified dataset on Narvik roads. The features here are:

- lat_ — latitude
- lon_ — longitude
- State_ — word description of road state (1: 'dry', 2: 'moist', 3: 'wet', 4: 'icy', 5: 'snowy', 6: 'slushy')
- Ta_mean, Ta_min, Ta_max — atmosphere temperature
- Tsurf_mean, Tsurf_min, Tsurf_max — surface temperature
- Water_mean, Water_min, Water_max — water layerw width (0 – 3 *mm*)
- Speed_mean, Speed_min, Speed_max — wind speed (in knots, 5 *knots* \approx 9.3 *km/h*)
- Height_mean, Height_min, Height_max — height of location above mean sea level
- Tdew_mean, Tdew_min, Tdew_max — dew point (*Celsius*)
- Friction_mean, Friction_min, Friction_max — friction value (0 – 1, 0 means no friction)
- Date, Time, date_time, FullDate — time and date
- Direction_min, Direction_max — wind direction (*degrees*)
- ClosestCity, location
- maxtempC, mintempC — day maximum and minimum of temperature (*Celsius*)
- totalSnow_cm — total snowfall (*cm*)
- sunHour — passed sun energy in *Sun – Hours* (A *Sun – Hour* is "1000 watts of energy shining on 1 square meter of surface for 1 hour")
- uvIndex — ultraviolet index
- moon_illumination — moon phase (*percents*)
- moonrise — time of Moon rise

- moonset — time of Moon set
- sunrise — time of Sun rise
- sunset — time of Sun set
- DewPointC — hourly dew point measurement (*Celsius*)
- FeelsLikeC — hourly Feels-like temperature (*Celsius*)
- HeatIndexC — hourly heat index (*Celsius*)
- WindChillC — hourly wind-chill index (*Celsius*)
- WindGustKmph — hourly wind gust measure (*km/h*)
- cloudcover — hourly cloud cover index (*percents*)
- humidity — hourly humidity (*percents*)
- precipMM — hourly precipitation (*mm*)
- pressure — hourly atmosphere pressure (*mbar*)
- tempC — hourly atmosphere temperature (*Celsius*)
- visibility — hourly visibility (0–10, 0 means poor visibility)
- winddirDegree — hourly wind direction (*degrees*)
- windspeedKmph — hourly wind speed (*km/h*)

2. Plotting a non-parametric estimation of PDF in form of a histogram and kernel density function for MRV (or probability law in case of discrete MRV)

Here one can observe histograms and KDE of density of Friction_mean and DewPointC variables as an example. The first looks like bimodal distribution and this bimodality is explained by State_: it is possible to separate Friction_mean with respect to $\text{State}_- < 4$, $\text{State}_- = 4$, $\text{State}_- > 4$. We use this information further in analysis.

Here visibility's distribution is also provided by both histogram and table. As one can see, the distribution is heavily focused on value 10.

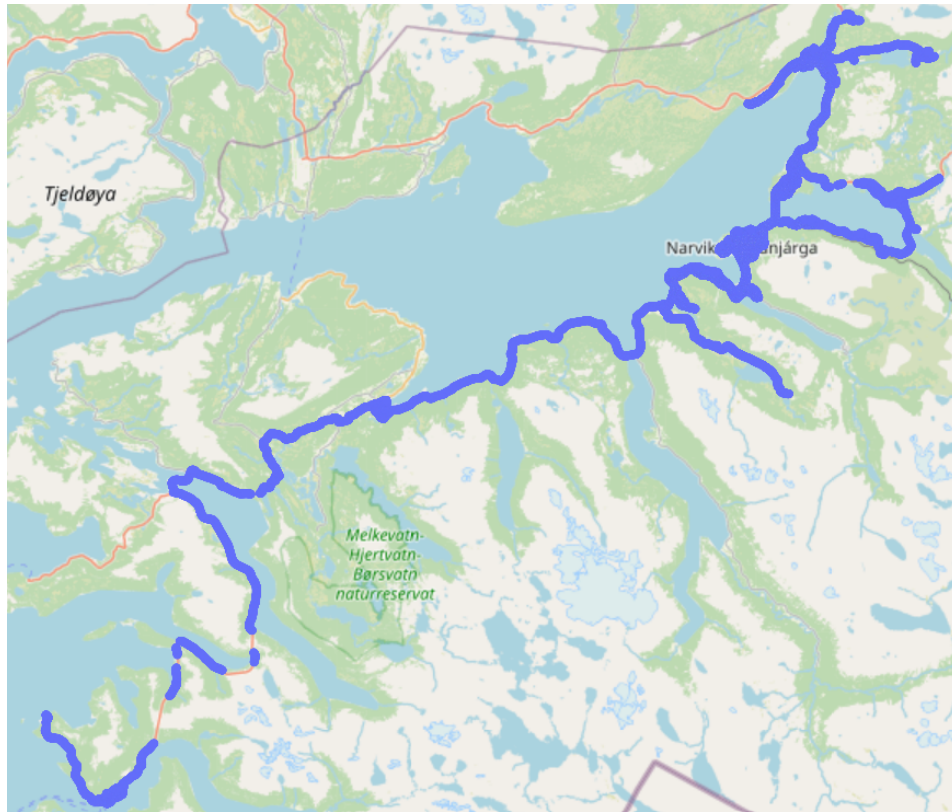


Figure 1: Geography of the data

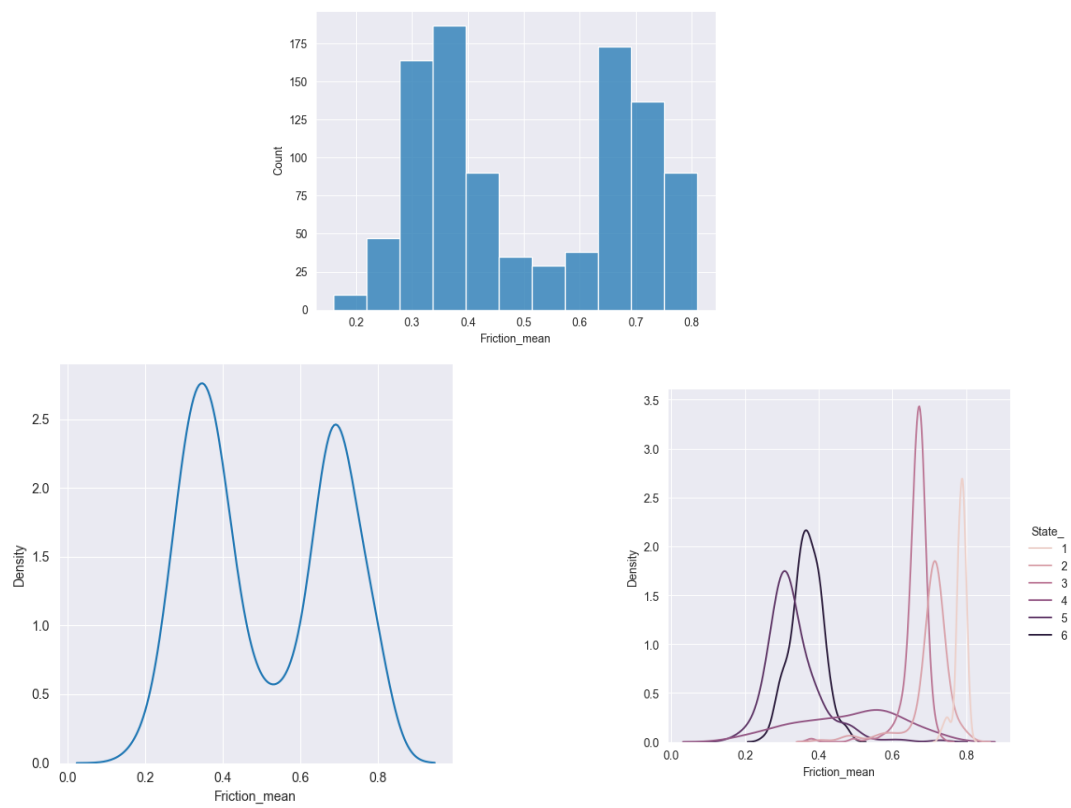


Figure 2: `Friction_mean`'s histogram, KDE and KDE conditional on `State_`

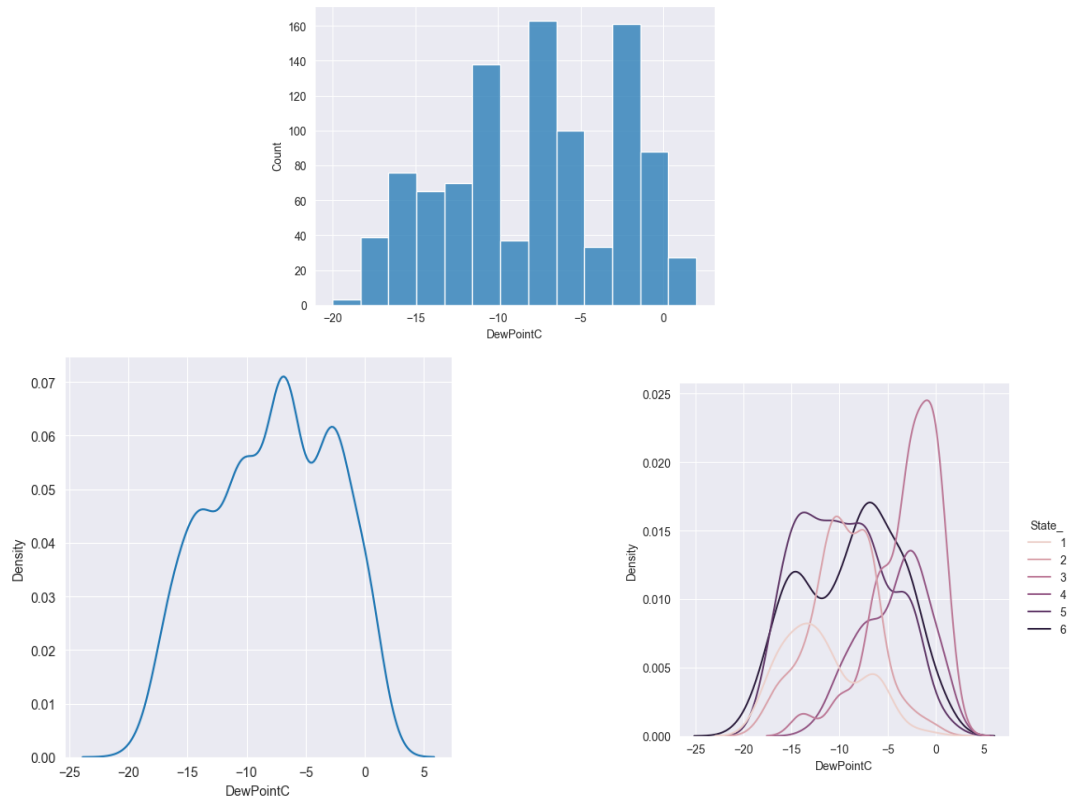


Figure 3: DewPointC's histogram, KDE and KDE conditional on State_

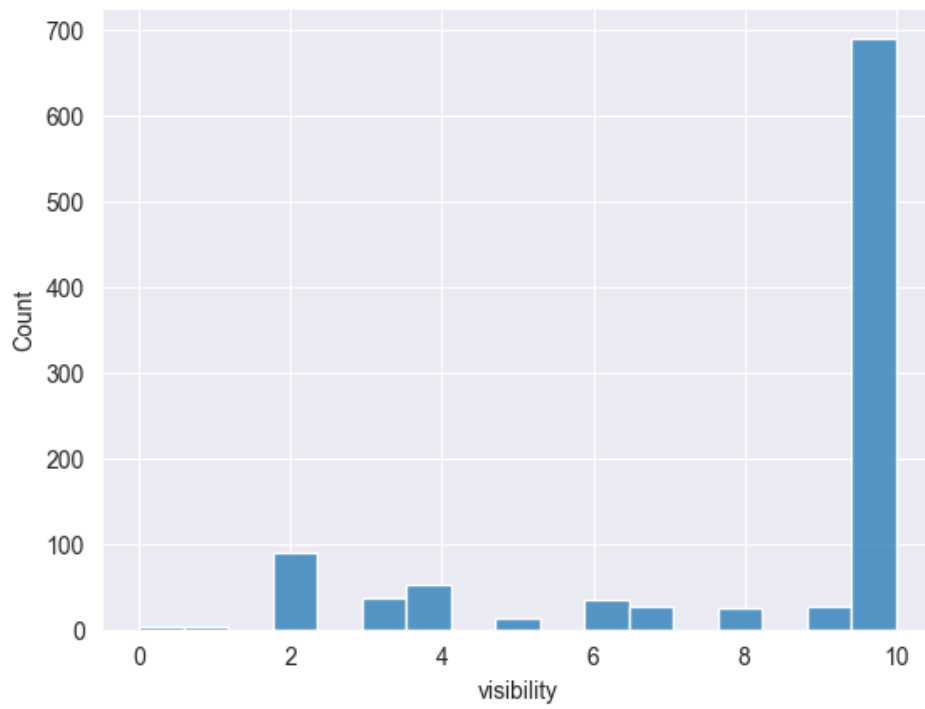


Figure 4: visibility's histogram

Moreover, pair-plot for a subset of features is provided in figure 5.

Since further we are interested in the construction of a model related to Friction_mean, it may

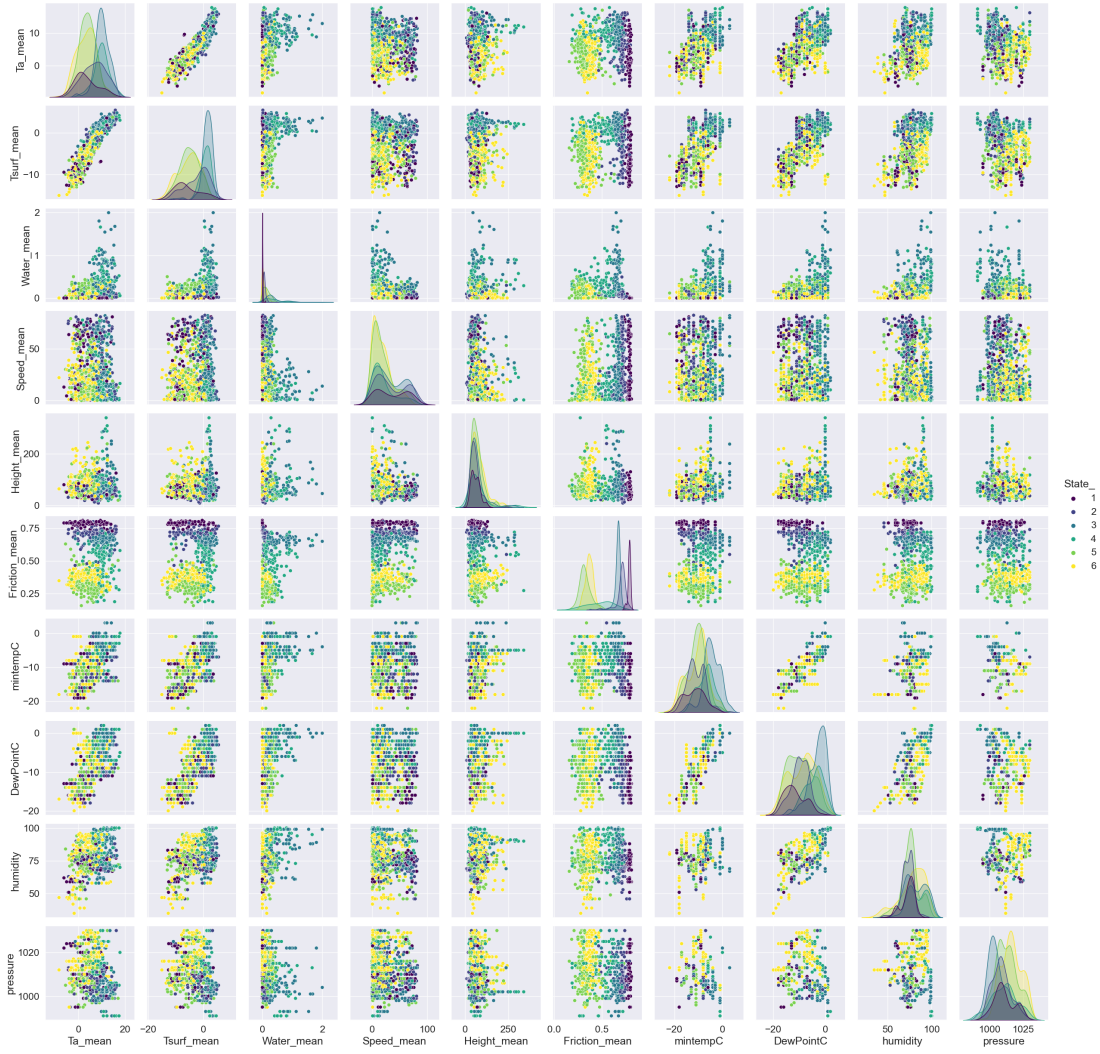


Figure 5: visibility's probability distribution

| visibility | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|
| p_i | 0.003 | 0.004 | 0.089 | 0.036 | 0.053 | 0.014 | 0.034 | 0.027 | 0.024 | 0.026 | 0.69 |

Table 1: visibility's distribution table

be useful to have a look at some joint distributions with several continuous features. In figures 6 and 7 KDE of joint distributions of Friction_mean with humidity and Height_mean correspondingly are presented.

3. Estimation of multivariate mathematical expectation and variance

Starting from now, we separate the dataset into three ones D_{l4} , $D_{=4}$ and $D_{>4}$ with respect to $\text{State_} < 4$, $\text{State_} = 4$, $\text{State_} > 4$ and emphasize our analysis only on D_{l4} with the following variables: Friction_mean, Ta_mean, Tsurf_mean, Water_mean, Speed_mean, Height_mean,

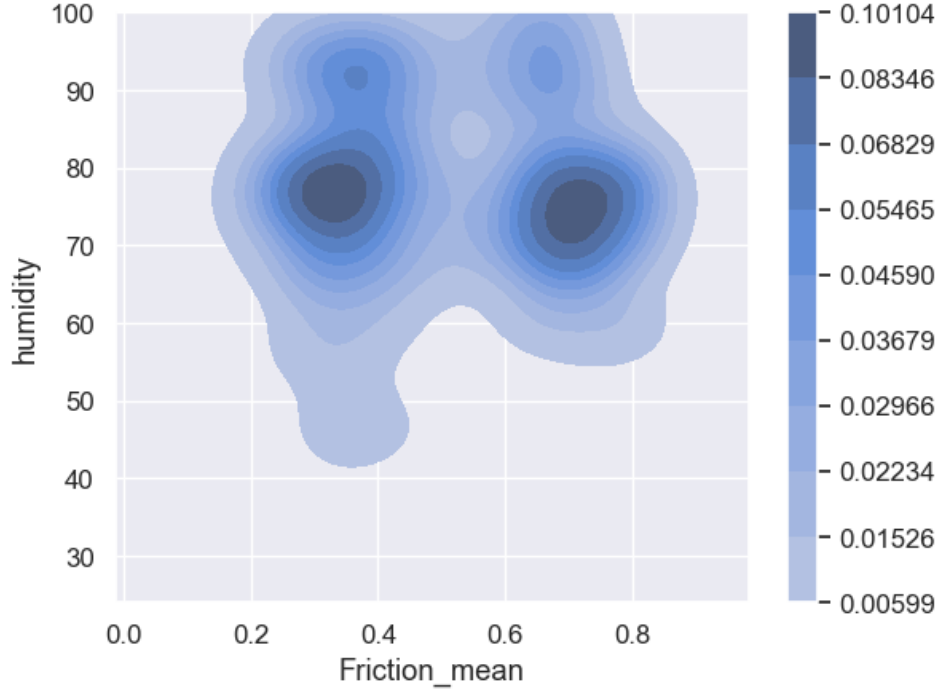


Figure 6: KDE of Friction_mean and humidity joint distribution density

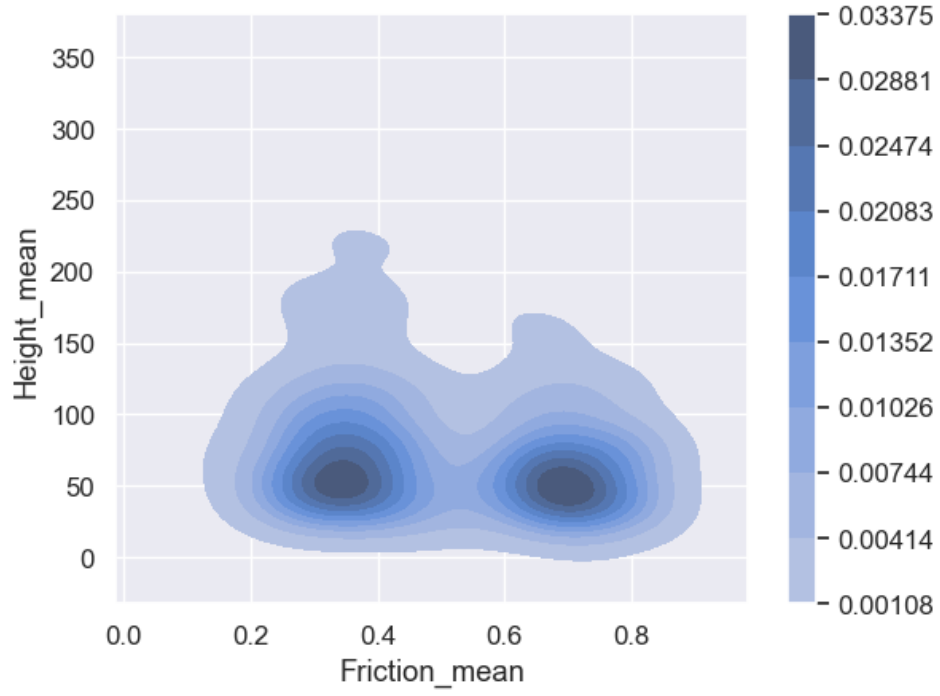


Figure 7: KDE of Friction_mean and Height_mean joint distribution density

maxtempC, mintempC, totalSnow_cm, sunHour, uvIndex, DewPointC, FeelsLikeC, HeatIndexC, WindChillC, WindGustKmph, cloudcover, humidity, precipMM, pressure, tempC, visibility, wind-speedKmph. In tables 2 we provide a part of mean vector and a part of covariance matrix estimate for D_{l4} .

| | | |
|--|----------------------|-----------|
| | <i>Friction_mean</i> | 0.702652 |
| | <i>Ta_mean</i> | 7.521634 |
| | <i>Tsurf_mean</i> | -1.397904 |
| | <i>Water_mean</i> | 0.210166 |
| | <i>Speed_mean</i> | 33.798909 |
| | <i>Height_mean</i> | 64.505240 |
| | <i>maxtempC</i> | -2.789346 |

| | | | | | | | |
|----------------------|----------------------|----------------|-------------------|-------------------|-------------------|--------------------|-----------------|
| | <i>Friction_mean</i> | <i>Ta_mean</i> | <i>Tsurf_mean</i> | <i>Water_mean</i> | <i>Speed_mean</i> | <i>Height_mean</i> | <i>maxtempC</i> |
| <i>Friction_mean</i> | 0.0039 | -0.1088 | -0.1120 | -0.0075 | 0.2216 | -0.2126 | -0.1273 |
| <i>Ta_mean</i> | -0.1088 | 24.2010 | 20.1588 | 0.5494 | -16.8366 | 27.5351 | 14.1173 |
| <i>Tsurf_mean</i> | -0.1120 | 20.1588 | 20.1056 | 0.5425 | -9.8468 | 20.8742 | 15.4006 |
| <i>Water_mean</i> | -0.0075 | 0.5494 | 0.5425 | 0.0901 | -1.7833 | 2.8684 | 0.6905 |
| <i>Speed_mean</i> | 0.2216 | -16.8366 | -9.8468 | -1.7833 | 682.6403 | -134.7463 | -3.7089 |
| <i>Height_mean</i> | -0.2126 | 27.5351 | 20.8742 | 2.8684 | -134.7463 | 1489.7925 | 32.6421 |
| <i>maxtempC</i> | -0.1273 | 14.1173 | 15.4006 | 0.6905 | -3.7089 | 32.6421 | 20.258915 |

Table 2: Parts of mean vector and covariance matrix for D_{l4}

4. Non-parametric estimation of conditional distributions, mathematical expectations and variances

In figure 8 KDEs of conditional distributions of Friction_mean given humidity's bins built on quartiles are illustrated on the left (for subdataset D_{l4}). At the same time on the right here are conditional mean (upper graph) and variance kernel estimates (lower graph). The points are colored with respect to State_.

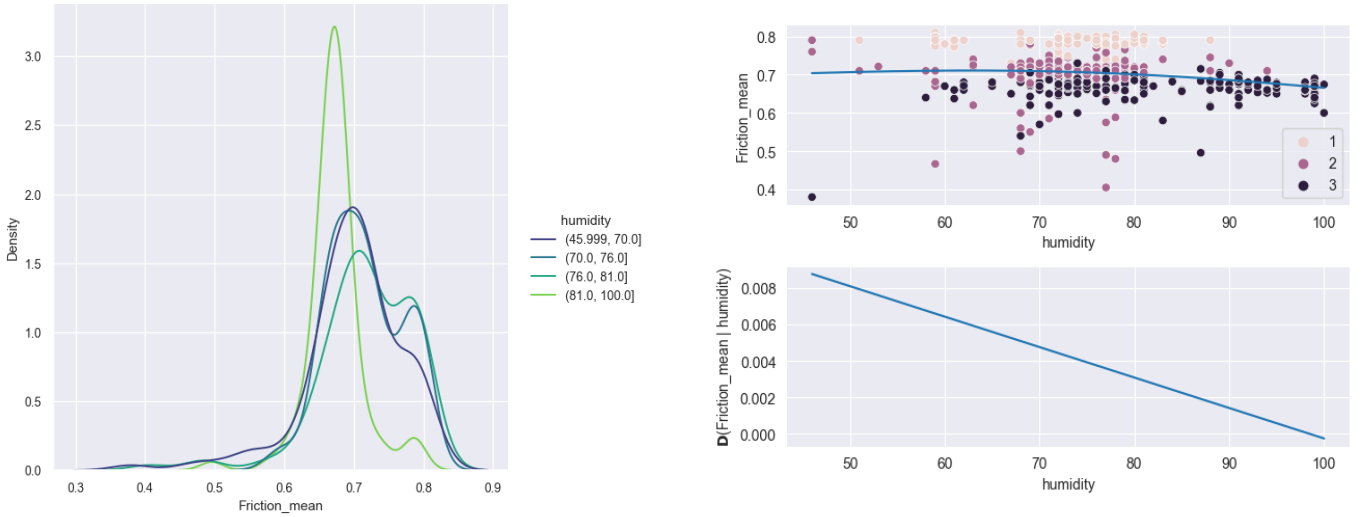


Figure 8: Conditional distributions KDEs, mean and variance of Friction_mean given humidity (or humidity's bins based on quartiles)

In addition, in figure 9 KDEs of conditional distributions of Friction_mean given Height_mean's bins built on quartiles are illustrated (for subdataset D_{b4}). On the right here are conditional mean (upper graph) and variance kernel estimates (lower graph). The points are colored with respect to State_.

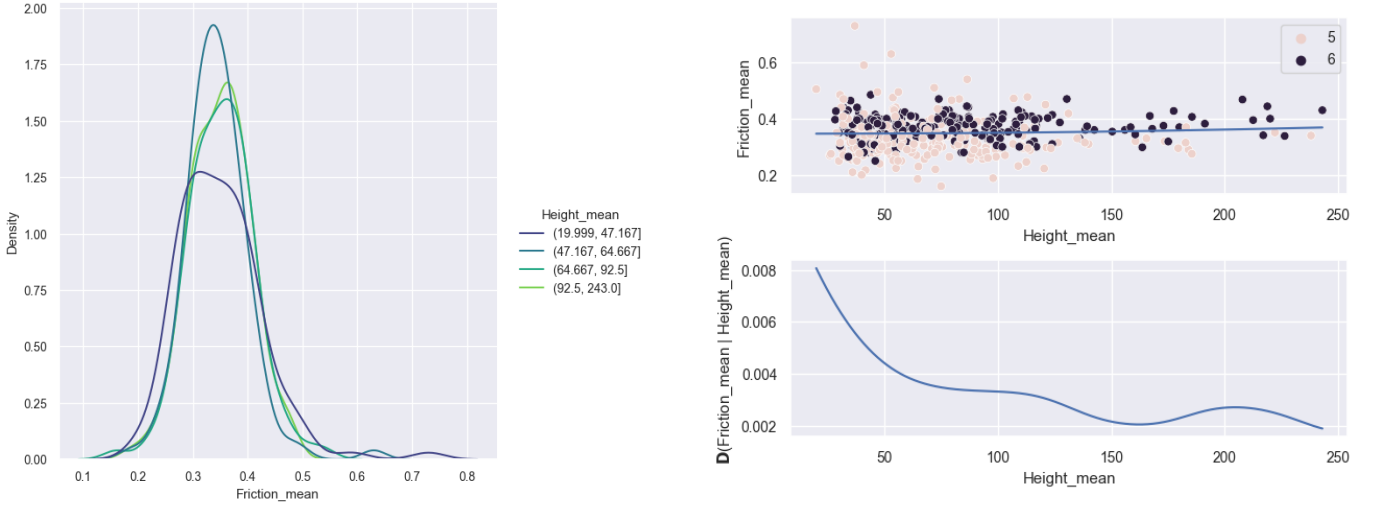


Figure 9: Conditional distributions KDEs, mean and variance of Friction_mean given Height_mean (or Height_mean's bins based on quartiles)

Conditional variance estimator is obtained from quite simple calculations. If $\mathbf{E}(\xi \mid \eta)$ is a regression of ξ on η , then $\mathbf{D}(\xi \mid \eta) = \mathbf{E}((\xi - \mathbf{E}(\xi \mid \eta))^2 \mid \eta)$ is a regression of $(\xi - \mathbf{E}(\xi \mid \eta))^2$ on η .

5. Estimation of pair correlation coefficients, confidence intervals for them and significance levels

5.1. Theory

Let $\hat{\rho}_n$ be estimator of $\rho = \rho(\xi, \eta)$ with $(\xi, \eta)^T \sim N(\mu, \Sigma)$. We test the hypothesis $H_0 : \rho = 0$. It is known that

$$t = \sqrt{n-2} \frac{\hat{\rho}_n}{\sqrt{1 - \hat{\rho}_n^2}} \sim t(n-2)$$

— a statistic which measures correlation. If our data do not obey Gaussian distribution, then the test utilizes critical values of asymptotic normal distribution $N(0, 1)$ (if n is large enough). It is possible to transform this statistic by z -transformation (Fisher):

$$z = \frac{1}{2} \ln \frac{1 + \hat{\rho}_n}{1 - \hat{\rho}_n}, \quad z_0 = \frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0}$$

which implies:

$$\sqrt{n-3}(z - z_0) \rightarrow^d N(0, 1)$$

Given significance level α , one can obtain both p -value and $1 - \alpha$ asymptotic confidence interval for z_0 . As soon as we derived the latter, we can find confidence interval for ρ by inverse of z -transformation:

$$z_l < z_0 < z_r \implies 1 - \frac{2}{e^{2z_l} + 1} < \rho < 1 - \frac{2}{e^{2z_r} + 1}.$$

5.2. Results

We conducted correlation significance analysis of Friction_mean versus other features chosen in section by the aforementioned test on significance level $\alpha = 0.05$. The results are provided in table 3.

| <i>Friction_mean vs</i> | $\hat{\rho}_n$ | ρ_{left} | ρ_{right} | $p - value$ |
|-------------------------|----------------|---------------|----------------|-------------|
| <i>DewPointC</i> | -0.4414 | -0.5159 | -0.3603 | 4e-21 |
| <i>FeelsLikeC</i> | -0.4310 | -0.5064 | -0.3490 | 4e-20 |
| <i>HeatIndexC</i> | -0.4484 | -0.5223 | -0.3678 | 8e-22 |
| <i>Height_mean</i> | -0.0886 | -0.1835 | 0.0080 | 0.0722 |
| <i>Speed_mean</i> | 0.1364 | 0.0404 | 0.2299 | 0.0055 |
| <i>Ta_mean</i> | -0.3555 | -0.4370 | -0.2682 | 9e-14 |
| <i>Tsurf_mean</i> | -0.4017 | -0.4796 | -0.3175 | 2e-17 |
| <i>Water_mean</i> | -0.4014 | -0.4793 | -0.3171 | 2e-17 |
| <i>WindChillC</i> | -0.4310 | -0.5064 | -0.3490 | 4e-20 |
| <i>WindGustKmph</i> | 0.1404 | 0.0445 | 0.2336 | 0.0043 |
| <i>cloudcover</i> | -0.3011 | -0.3864 | -0.2107 | 4e-10 |
| <i>humidity</i> | -0.1659 | -0.2582 | -0.0705 | 7e-04 |
| <i>maxtempC</i> | -0.4546 | -0.5279 | -0.3743 | 2e-22 |
| <i>mintempC</i> | -0.3872 | -0.4663 | -0.3020 | 3e-16 |
| <i>precipMM</i> | -0.2133 | -0.3035 | -0.1192 | 1e-05 |
| <i>pressure</i> | 0.1204 | 0.0242 | 0.2144 | 0.0144 |
| <i>sunHour</i> | 0.0601 | -0.0366 | 0.1557 | 0.2227 |
| <i>tempC</i> | -0.4489 | -0.5227 | -0.3683 | 7e-22 |
| <i>totalSnow_cm</i> | -0.1655 | -0.2579 | -0.0701 | 7e-04 |
| <i>uvIndex</i> | 0.2073 | 0.1131 | 0.2978 | 2e-05 |
| <i>visibility</i> | 0.2561 | 0.1637 | 0.3441 | 1e-07 |
| <i>windspeedKmph</i> | 0.1332 | 0.0371 | 0.2267 | 0.0067 |

Table 3: Correlation significance of Friction_mean vs others test results

As one can see, there is no significance of correlation with sunHour and Height_mean ($p - value > \alpha$) so we omit these features further and use others to build first linear regression model.

6. Task formulation for regression

Our purpose is to build regression model for Friction_mean on data D_{I4} by features Ta_mean, Tsurf_mean, Water_mean, Speed_mean, maxtempC, mintempC, totalSnow_cm, uvIndex, DewPointC, FeelsLikeC, HeatIndexC, WindChillC, WindGustKmph, cloudcover, humidity, precipMM, pressure, tempC, visibility, windspeedKmph with further selection of predictors and, if any, deleting some observations by means of different statistical tools.

7. Regression model construction

7.1. Model 1

The first model was built with the aforementioned variables. To check its quality, we look at R_{adj}^2 and MSE scores as well as the residuals of the model: their normal probability plot (QQ-plot versus standard normal distribution after standardization).

This model has $R_{adj}^2 \approx 0.233$ and $MSE \approx 0.0028$. The QQ-plot of the residuals is illustrated in figure 10. By no means can one treat these residuals as normally distributed so this model was not accepted.

7.2. Multicollinearity and partial correlation analysis

In order to improve model quality, we decided to look at correlation between our predictors to remove all extra variables which correlate significantly with some others. The heatmap for correlation coefficients between our predictors is demonstrated in 11. As has been seen, the whole group of 'temperature' features has strong correlation. We removed all of these variables except Tsurf_mean as surface temperature closely relates to roads. Furthermore, feature WindGustKmph heavily correlates with windspeedKmph so we also removed the former further. However, the latter is also extra since it replicates feature Speed_mean but in hourly manner so it is quite discrete. At last, variable visibility is discrete (has 11 unique values) and its distribution is heavily focused on value 10, thus, we consider interaction with this feature difficult in building regression model and removed it.

Also we analyzed partial correlations between our target variable Friction_mean and others to find potential predictors which correlates with target even if the influence of other variables is neutralized.

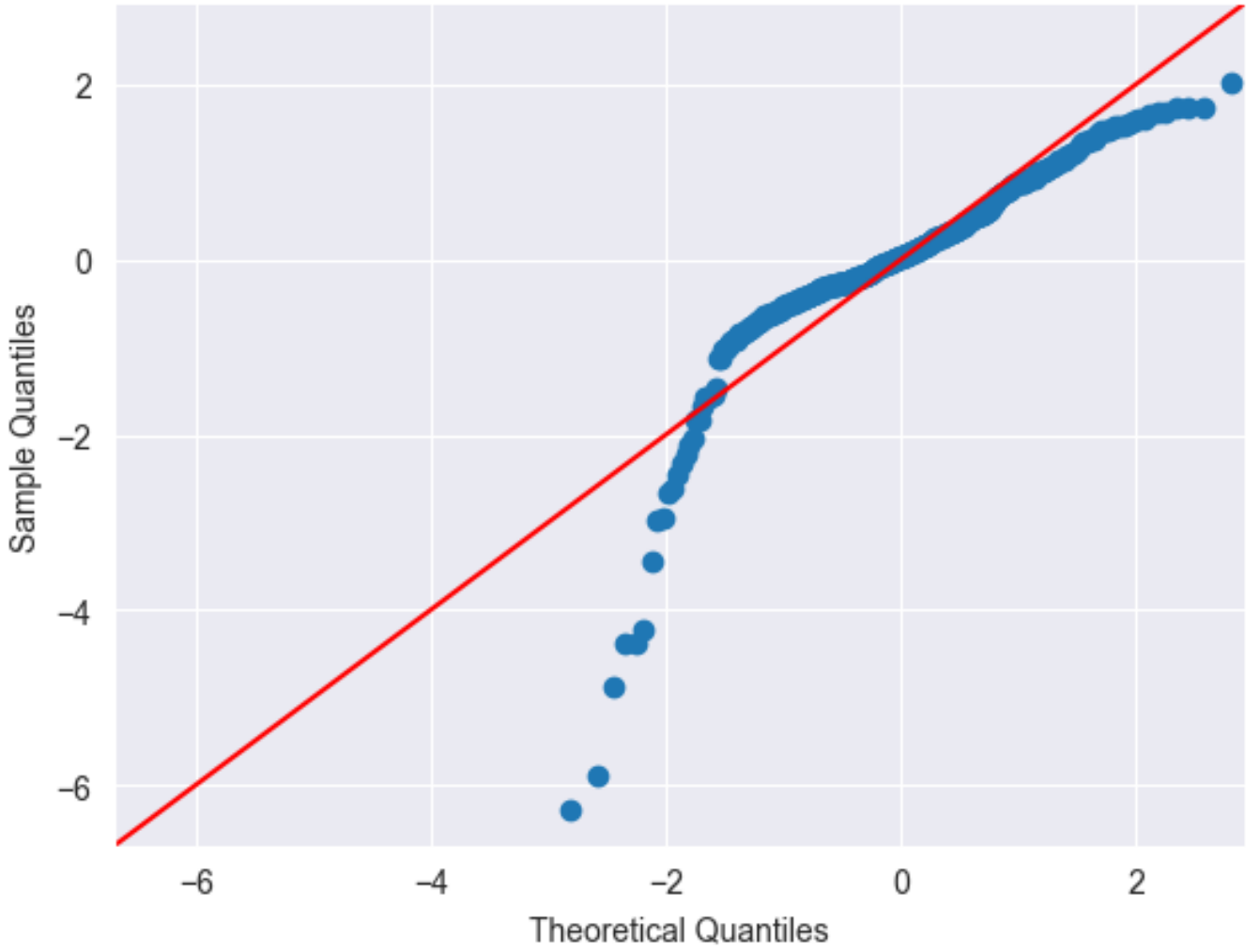


Figure 10: Normal probability plot for the residuals of Model 1

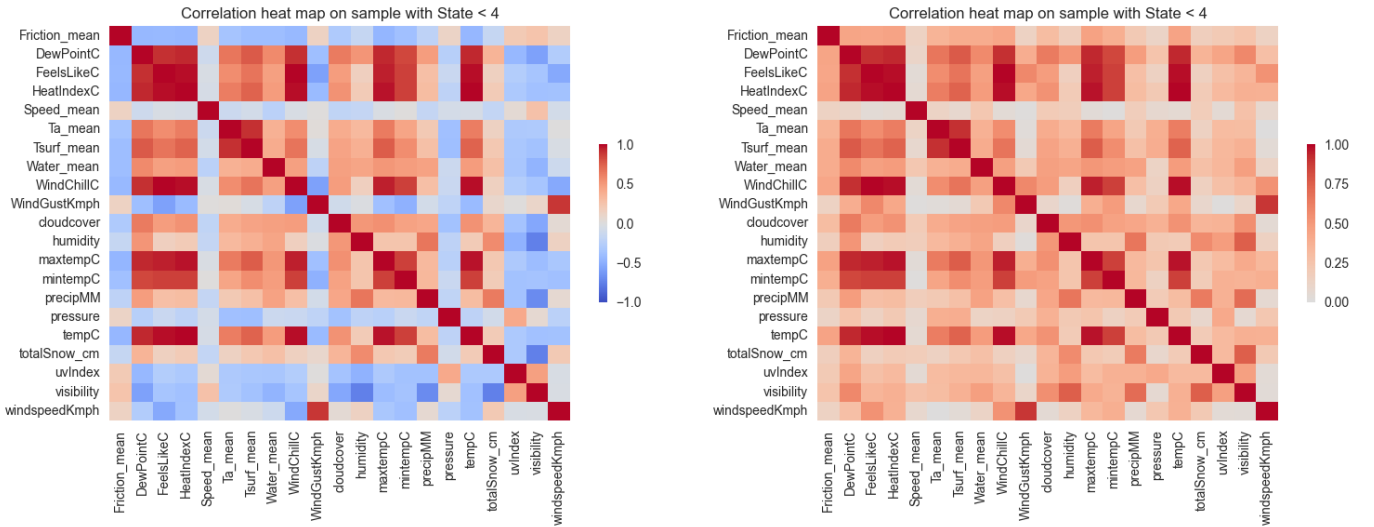


Figure 11: Heatmap for correlation coefficients: one is colored depending on the sign of the correlation, another is colored by absolute values of correlation coefficients

Partial correlation measures correlation of two variables if influence of other variables is removed:

$$\begin{aligned}\rho(\xi, \eta \mid \alpha_1, \dots, \alpha_n) &= \rho(\xi - \tilde{\xi}, \eta - \tilde{\eta}) \\ \tilde{\xi} &= \arg \min_{\xi^* \in K} \mathbf{E}(\xi - \xi^*)^2 \\ \tilde{\eta} &= \arg \min_{\eta^* \in K} \mathbf{E}(\eta - \eta^*)^2 \\ K &= \{a_0 + a_1\alpha_1 + \dots + a_n\alpha_n\}\end{aligned}$$

In figure 12 one can see partial correlation heatmap. It is evident that no variable has own separate influence on the target so we have to deal with all variables selected above.

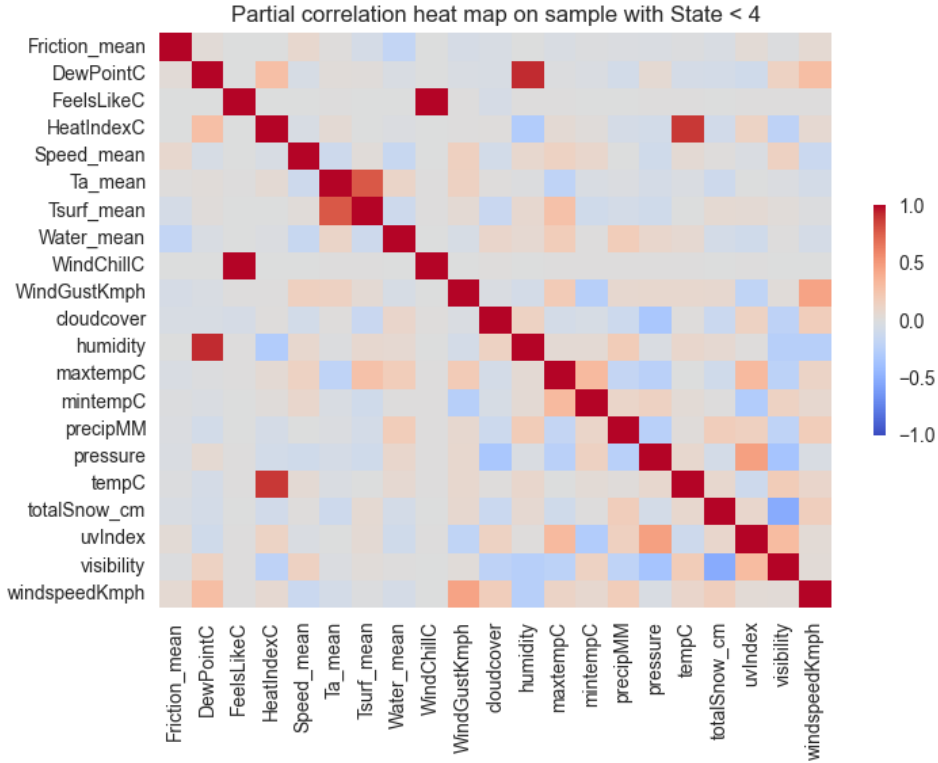


Figure 12: Partial correlation heatmap

As a result of variables selection, we further work with the following variables: Tsurf_mean, Water_mean, Speed_mean, totalSnow_cm, uvIndex, DewPointC, cloudcover, humidity, precipMM, pressure.

7.3. Outlier with respect to distribution removal

Besides, we decided to remove outliers using histogram of the target variable and Mahalanobis distance on the regressors. From graph 13 one can conclude that observations with Friction_mean < 0.5 may be treated as outliers since at least they heavily influence on regression.

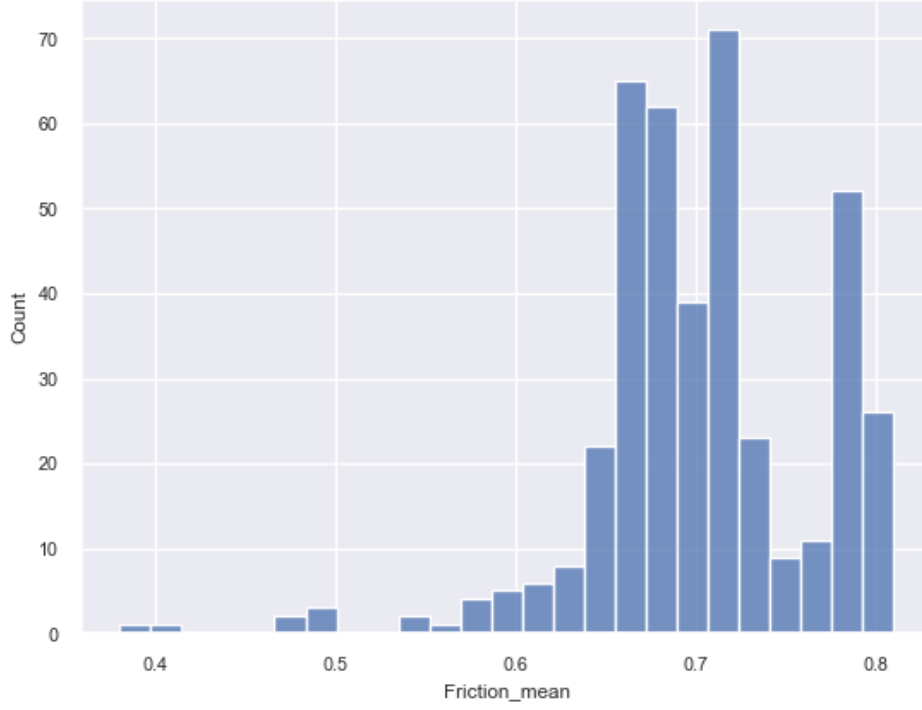


Figure 13: Friction_mean histogram, State_ < 4

Mahalanobis' distance

Suppose S is a positive definite $d \times d$ matrix. The Mahalanobis' distance between two points $x, y \in \mathbf{R}^d$ with these matrix is defined as follows:

$$d_M^2(x, y; S) = (x - y)S^{-1}(x - y)^T.$$

If $\xi \sim N(\mu, \Sigma)$, then $d^2(\xi, \mu; \Sigma) \sim \chi^2(d)$. Using these information one can construct statistical test for outliers detection if data are normal, otherwise, one can just analyze the graph of Mahalanobis' distribution of all observations with respect to empirical distribution.

As our data are not normal, we analyzed the Mahalanobis' distance graph 14. We treat the observations with $d_M > 8$ as explicit outliers so we remove them in further analysis as well as the observations with Friction_mean < 0.5.

7.4. Model 2

After multicollinearity analysis and distribution outliers analysis, we trained linear regression model with aforementioned features. Its $R_{adj}^2 \approx 0.408$ and $MSE = 0.0017$ which are significantly better than of the Model 1. However, the residuals' distribution is far from being normal as shown in the QQ-plot 15

Consider a significance test of Model 2 regression coefficients. Its result on significance level $\alpha = 0.05$ are presented in table 4 .

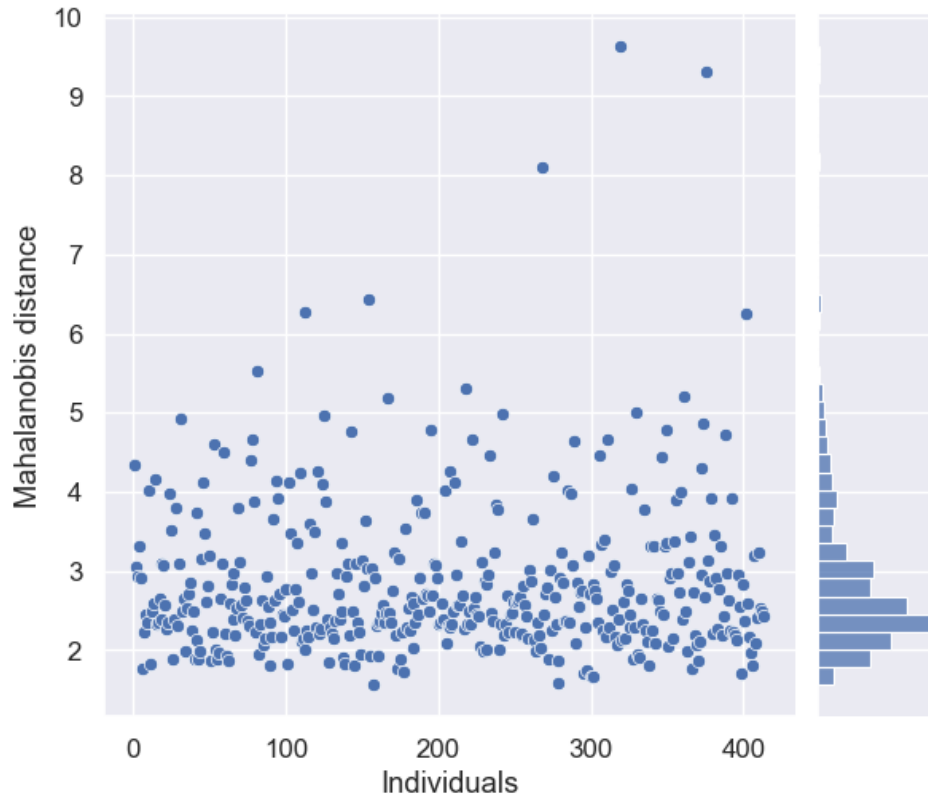


Figure 14: Mahalanobis' distance jointplot

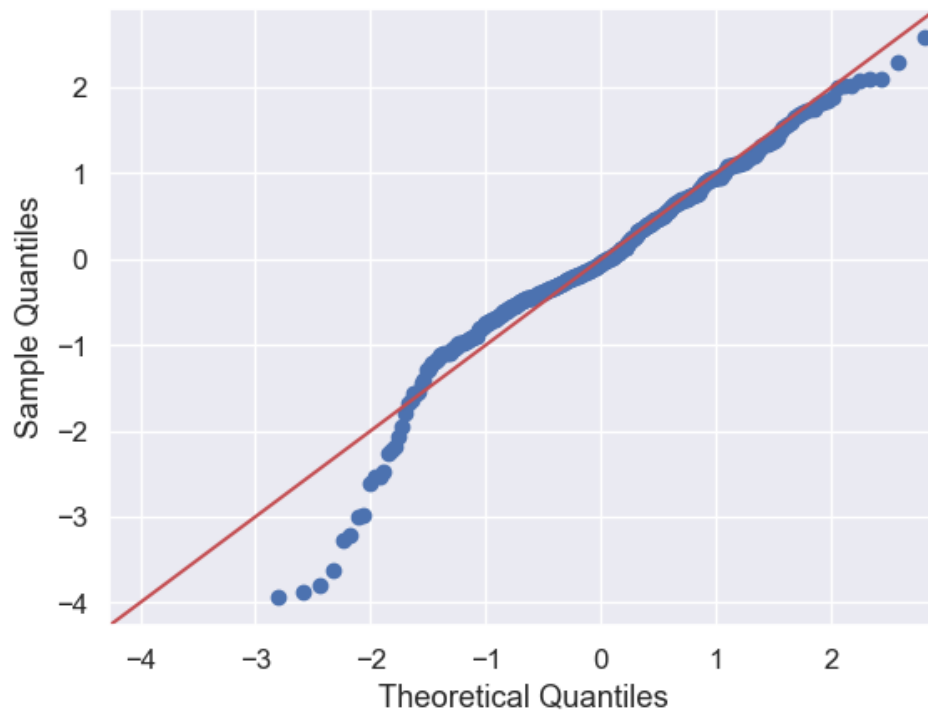


Figure 15: Normal probability plot for the residuals of Model 2

One can infer that `Speed_mean`, `totalSnow_cm`, `uvindex`, `precipMM` and `pressure` have no evidence to be proved significant so we decided to remove these feature. In spite of $p - value$ of the

| | <i>coef</i> | <i>std err</i> | <i>t</i> | $P > t $ | [0.025 | 0.975] |
|---------------------|-------------|----------------|----------|-------------|--------------|-------------|
| <i>intercept</i> | 0.6627 | 0.33 | 2.007 | 0.045 | 0.014 | 1.312 |
| <i>Tsurf_mean</i> | −0.0044 | 0.001 | −7.900 | ≈ 0 | −0.006 | −0.003 |
| <i>Water_mean</i> | −0.0641 | 0.009 | −6.952 | ≈ 0 | −0.082 | −0.046 |
| <i>Speed_mean</i> | 3e-06 | 8e-05 | 0.037 | 0.971 | ≈ -0 | ≈ 0 |
| <i>totalSnow_cm</i> | −0.0005 | 0.001 | −0.821 | 0.412 | −0.002 | 0.001 |
| <i>windex</i> | 0.0093 | 0.005 | 1.819 | 0.070 | −0.001 | 0.019 |
| <i>cloudcover</i> | −0.0002 | 9e-05 | −1.939 | 0.053 | ≈ -0 | 2e-06 |
| <i>humidity</i> | 0.0009 | ≈ 0 | 3.002 | 0.003 | ≈ 0 | 0.002 |
| <i>precipMM</i> | −0.0081 | 0.010 | −0.788 | 0.431 | −0.0028 | 0.012 |
| <i>pressure</i> | −2e-05 | ≈ 0 | −0.064 | 0.949 | −0.001 | 0.001 |

Table 4: Regression coefficients significance test results: coefficients, their standard error, test statistic value, p-values and 95% confidence intervals

coefficient of cloudcover being a bit above 0.05, we did not remove it in further analysis.

So, the left features are Tsurf_mean, Water_mean, cloudcover and humidity (and intercept!).

7.5. Outliers with respect to regression analysis

Using the aforementioned features, we analyzed outliers with respect to regression (i.e. observations with high leverage).

Cook's distance

Let $y = Xw + \varepsilon$, $w \in \mathbb{R}^{p+1}$ and $\varepsilon \sim N(0, \sigma^2 I)$. It is evident that $\hat{w} = \hat{w}_{OLS} = (X^T X)^{-1} X^T y$ and $\hat{y} = X(X^T X)^{-1} X^T y = Hy$.

Residual for y_i : $r_i = y_i - \hat{y}_i$, deleted residual: $r_i^{(i)} = y_i - \hat{y}_i^{(i)}$, where $\hat{y}_i^{(i)}$ — estimator of y_i with regression built on data without i -th individual.

It follows from the definition that $\mathbb{D}r_i = \sigma^2(1 - h_{ii})$ and then $\frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ — studentized residual, $\hat{\sigma}^2 = \frac{1}{n-p} \sum_{k=1}^n (y_k - \hat{y}_k)^2$.

Cook's distance from i -th individual to regression is:

$$D_i = \frac{\sum_{k=1}^n (\hat{y}_k - \hat{y}_k^{(i)})^2}{p\hat{\sigma}^2} = \frac{(y_i - \hat{y}_i)^2}{p\hat{\sigma}^2} \frac{h_{ii}}{(1 - h_{ii})^2}$$

There is a rule-of-thumb: the observations with $D_i > \frac{4}{n}$ are believed to be outliers [1].

In our case, the graph of Cook's distance is demonstrated in figure 16 where the threshold is approximately equal to 0.01. We treat these observations as outliers since they have significant leverage for regression construction.

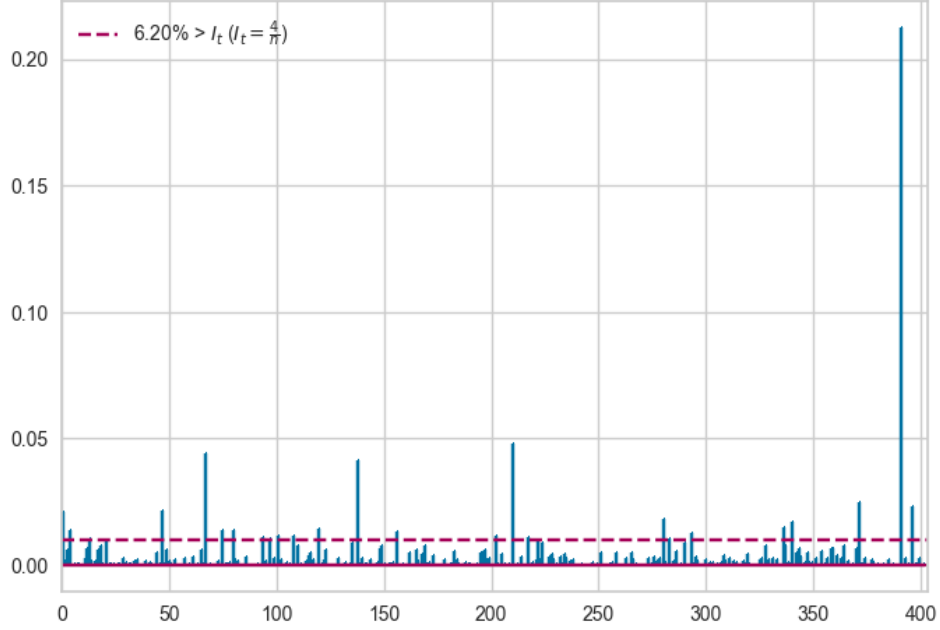


Figure 16: Cook's distance plot

7.6. Model 3

The model built on the data with aforementioned features has $R_{adj}^2 \approx 0.565$ and $MSE \approx 0.001$. These values are better than those of Model 2. If one looks at the QQ-plot (fig. 17) of the regression residuals, they can treat the residuals as practically normal.

We consider this model as final one. Its description in terms of coefficients significance are listed in table 5.

| | <i>coef</i> | <i>std err</i> | <i>t</i> | $P > t $ | [0.025 | 0.975] |
|-------------------|-------------|----------------|----------|-------------|------------------|----------|
| <i>intercept</i> | 0.7001 | 0.014 | 48.744 | ≈ 0 | 0.672 | 0.728 |
| <i>Tsurf_mean</i> | -0.0054 | ≈ 0 | -10.965 | ≈ 0 | -0.006 | -0.004 |
| <i>Water_mean</i> | -0.0854 | 0.008 | -10.136 | ≈ 0 | -0.102 | -0.069 |
| <i>cloudcover</i> | -0.0001 | 7e-05 | -2.017 | 0.044 | ≈ -0 | -3.6e-06 |
| <i>humidity</i> | 0.0004 | ≈ 0 | 1.826 | 0.069 | $\approx -3e-05$ | 0.001 |

Table 5: Regression coefficients significance test results on Model 3: coefficients, their standard error, test statistic value, p-values and 95% confidence intervals

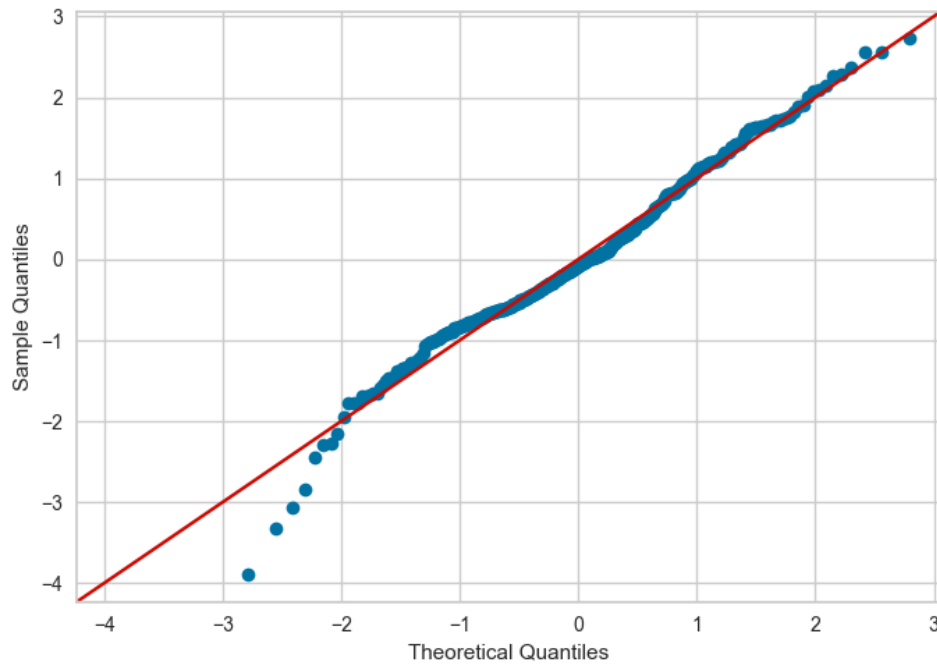


Figure 17: Normal probability plot for the residuals of Model 3

8. Appendix

The Python notebook related to the aforementioned calculations is presented in Github [2].

Bibliography

1. Fox J. Regression Diagnostics. — SAGE Publications, Inc., 1991. — Access mode: <https://methods.sagepub.com/book/regression-diagnostics>.
2. Grigorev D. Code repository. — <https://github.com/dmitry-grigorev/MultivarAnalysis/blob/master/Lab1/Lab2notebook.ipynb>. — 2022.