1. Explore the properties of the data, including
   1. Average of each variable

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **FA** | **VA** | **CA** | **RS** | **Ch** | **FSD** | **TSD** | **Density** | **pH** | **Sulphates** | **Alc** |
| 8.32 | 0.53 | 0.27 | 2.54 | 0.09 | 15.88 | 46.47 | 1.00 | 3.31 | 0.66 | 10.42 |

* 1. Histogram of the quality score variable

Chart, histogram

Description automatically generated

* 1. Relationship between the physicochemical properties and the wine quality score using the correlation function.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **FA** | **VA** | **CA** | **RS** | **Ch** | **FSD** | **TSD** | **Density** | **pH** | **Sulphates** | **Alc** |
| **Correlation with QS** | 0.124 | -0.391 | 0.226 | 0.014 | -0.129 | -0.051 | -0.185 | -0.175 | -0.058 | 0.251 | 0.476 |

1. Split the data into training and test sets. Use random sampling, with 80%

allocated to the training set and 20% allocated to the test set. Use the linear regression to build a predictive model for the wine quality score using the training set data. Explore the following:

* 1. Multiple linear regression model with all physicochemical properties.

|  |
| --- |
| Call:  lm(formula = QS ~ ., data = trainData)  Residuals:  Min 1Q Median 3Q Max  -2.65050 -0.38391 -0.04515 0.45702 2.01409  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.490e+01 2.405e+01 0.620 0.53569  FA 1.242e-02 2.998e-02 0.414 0.67877  VA -1.021e+00 1.398e-01 -7.304 4.91e-13 \*\*\*  CA -1.296e-01 1.687e-01 -0.768 0.44262  RS 6.203e-03 1.749e-02 0.355 0.72284  Ch -2.059e+00 4.711e-01 -4.371 1.34e-05 \*\*\*  FSD 4.145e-03 2.502e-03 1.656 0.09791 .  TSD -3.561e-03 8.351e-04 -4.264 2.16e-05 \*\*\*  Density -1.024e+01 2.455e+01 -0.417 0.67667  pH -5.783e-01 2.193e-01 -2.637 0.00846 \*\*  Sulphates 8.651e-01 1.329e-01 6.508 1.09e-10 \*\*\*  Alc 2.908e-01 3.040e-02 9.564 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.6634 on 1267 degrees of freedom  Multiple R-squared: 0.3551, Adjusted R-squared: 0.3495  F-statistic: 63.41 on 11 and 1267 DF, p-value: < 2.2e-16 |

* 1. Stepwise regression.

|  |
| --- |
| Call:  lm(formula = QS ~ VA + Ch + FSD + TSD + pH + Sulphates + Alc,  data = trainData)  Residuals:  Min 1Q Median 3Q Max  -2.65513 -0.37980 -0.04544 0.46399 2.02174  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 4.7312900 0.4552581 10.393 < 2e-16 \*\*\*  VA -0.9654987 0.1151049 -8.388 < 2e-16 \*\*\*  Ch -2.1511320 0.4473199 -4.809 1.70e-06 \*\*\*  FSD 0.0045550 0.0024584 1.853 0.0641 .  TSD -0.0037041 0.0007855 -4.715 2.68e-06 \*\*\*  pH -0.5871845 0.1336715 -4.393 1.21e-05 \*\*\*  Sulphates 0.8437438 0.1278251 6.601 5.99e-11 \*\*\*  Alc 0.2972283 0.0189738 15.665 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.6626 on 1271 degrees of freedom  Multiple R-squared: 0.3546, Adjusted R-squared: 0.3511  F-statistic: 99.77 on 7 and 1271 DF, p-value: < 2.2e-16 |

* 1. Any potential nonlinear relationships.

Linear regression with quadratic and logarithm terms added for 7 statically significant predictors identified by the previous stepwise: VA, Ch, FSD, TSD, pH, Sulphates, Alc

|  |
| --- |
| Call:  lm(formula = QS ~ VA + I(VA^2) + I(log(VA)) + Ch + I(Ch^2) +  I(log(Ch)) + TSD + I(TSD^2) + I(log(TSD)) + FSD + I(FSD^2) +  I(log(FSD)) + pH + Sulphates + I(Sulphates^2) + I(log(Sulphates)) +  Alc, data = trainData)  Residuals:  Min 1Q Median 3Q Max  -2.69813 -0.38174 -0.03291 0.43054 2.02515  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7.372e+00 2.572e+00 2.866 0.004224 \*\*  VA 2.278e+00 1.967e+00 1.158 0.247177  I(VA^2) -1.418e+00 8.460e-01 -1.676 0.094036 .  I(log(VA)) -7.178e-01 5.291e-01 -1.357 0.175103  Ch 7.050e+00 3.620e+00 1.948 0.051662 .  I(Ch^2) -9.158e+00 4.714e+00 -1.943 0.052239 .  I(log(Ch)) -7.194e-01 2.624e-01 -2.741 0.006210 \*\*  TSD -1.260e-02 4.270e-03 -2.952 0.003218 \*\*  I(TSD^2) 3.716e-05 1.420e-05 2.616 0.008995 \*\*  I(log(TSD)) 1.776e-01 1.315e-01 1.351 0.177011  FSD -4.246e-03 1.737e-02 -0.245 0.806872  I(FSD^2) 3.004e-05 2.133e-04 0.141 0.888014  I(log(FSD)) 1.334e-01 1.455e-01 0.917 0.359365  pH -7.846e-01 1.348e-01 -5.820 7.45e-09 \*\*\*  Sulphates -6.912e+00 3.181e+00 -2.173 0.029958 \*  I(Sulphates^2) 1.071e+00 8.354e-01 1.282 0.199905  I(log(Sulphates)) 4.542e+00 1.373e+00 3.308 0.000966 \*\*\*  Alc 2.778e-01 1.958e-02 14.184 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.6469 on 1261 degrees of freedom  Multiple R-squared: 0.3896, Adjusted R-squared: 0.3814  F-statistic: 47.35 on 17 and 1261 DF, p-value: < 2.2e-16 |

Running stepwise on the previous polynomial model

|  |
| --- |
| Call:  lm(formula = QS ~ I(VA^2) + Ch + I(Ch^2) + I(log(Ch)) + TSD +  I(TSD^2) + I(log(TSD)) + I(log(FSD)) + pH + Sulphates + I(log(Sulphates)) +  Alc, data = trainData)  Residuals:  Min 1Q Median 3Q Max  -2.74200 -0.38463 -0.03049 0.43293 1.94216  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 6.146e+00 1.163e+00 5.283 1.50e-07 \*\*\*  I(VA^2) -6.305e-01 9.359e-02 -6.736 2.46e-11 \*\*\*  Ch 5.977e+00 3.449e+00 1.733 0.08332 .  I(Ch^2) -7.837e+00 4.530e+00 -1.730 0.08389 .  I(log(Ch)) -6.528e-01 2.498e-01 -2.614 0.00906 \*\*  TSD -1.316e-02 3.992e-03 -3.298 0.00100 \*\*  I(TSD^2) 3.884e-05 1.382e-05 2.811 0.00502 \*\*  I(log(TSD)) 1.925e-01 1.204e-01 1.600 0.10994  I(log(FSD)) 8.972e-02 4.451e-02 2.016 0.04403 \*  pH -7.719e-01 1.333e-01 -5.791 8.81e-09 \*\*\*  Sulphates -2.895e+00 5.623e-01 -5.149 3.04e-07 \*\*\*  I(log(Sulphates)) 2.869e+00 4.184e-01 6.857 1.10e-11 \*\*\*  Alc 2.796e-01 1.939e-02 14.420 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.6467 on 1266 degrees of freedom  Multiple R-squared: 0.3877, Adjusted R-squared: 0.3819  F-statistic: 66.79 on 12 and 1266 DF, p-value: < 2.2e-16 |

1. Obtain the predictions for wine quality score for the test set data. Compute the predictive accuracy metrics for the test set. Comment on which model in your consideration set performs best.
2. Investigate the best performing predictive model you have chosen. Discuss what your chosen model reveals about the relationships between the wine quality score and the physicochemical properties.
3. What are the limitations of your analysis? Discuss.

Produce a 3-page report that summarizes your analysis