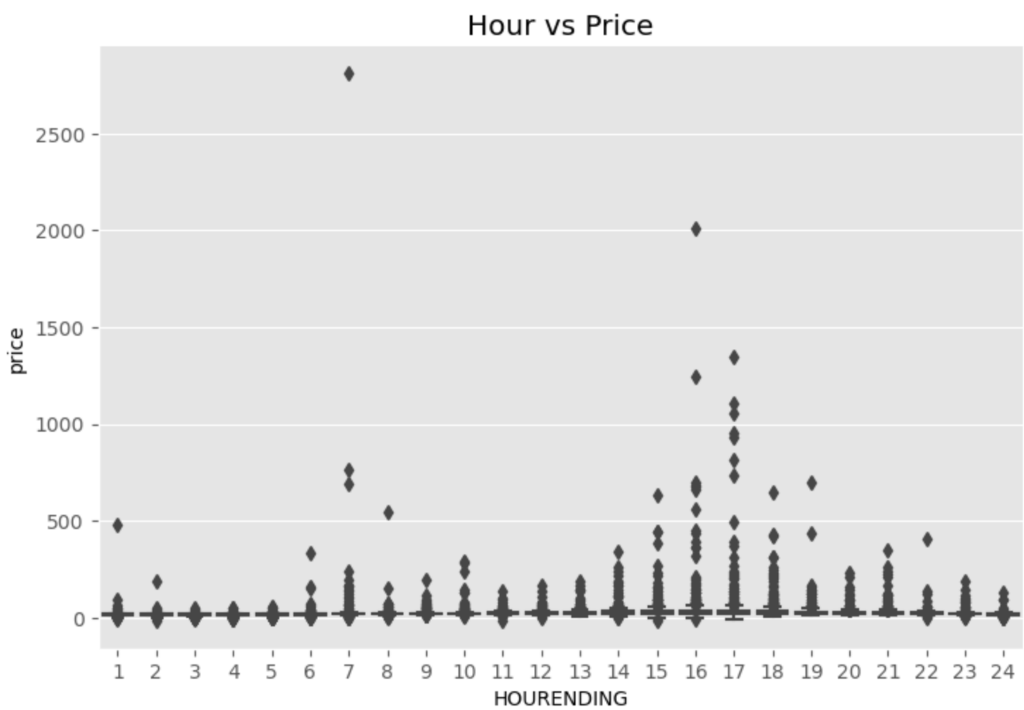
Q3 report

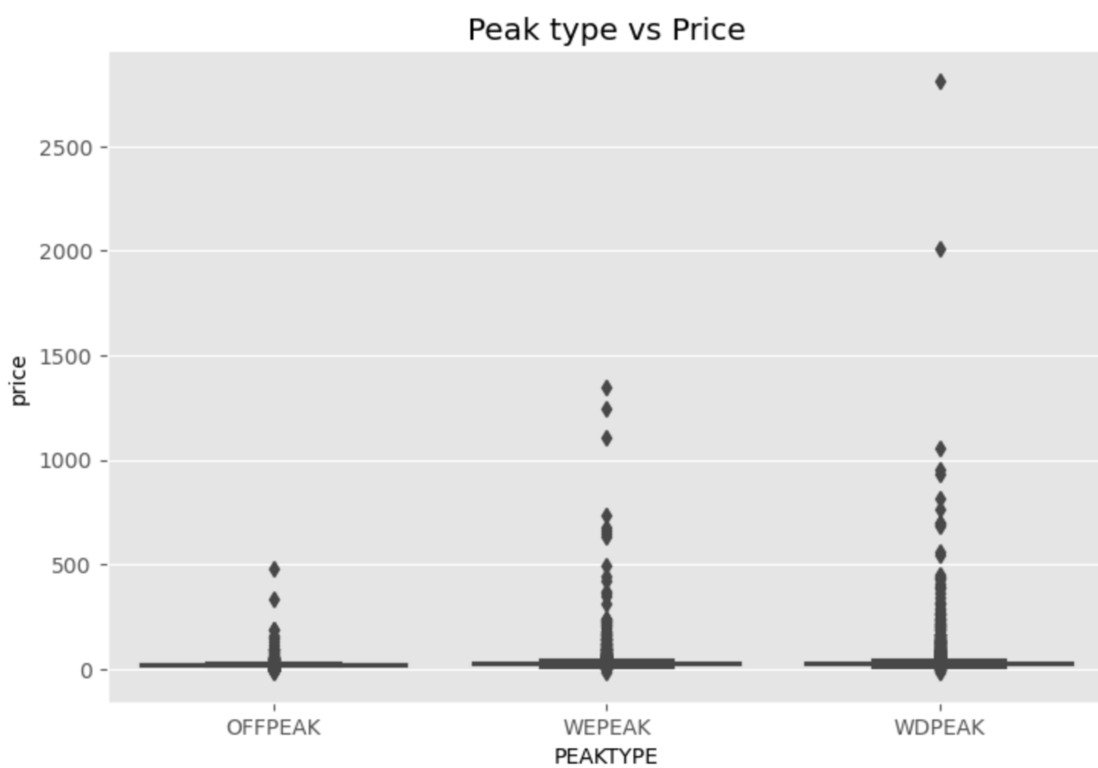
This question is slightly more exploratory than the first two, so I did my best analysis based on my knowledge.

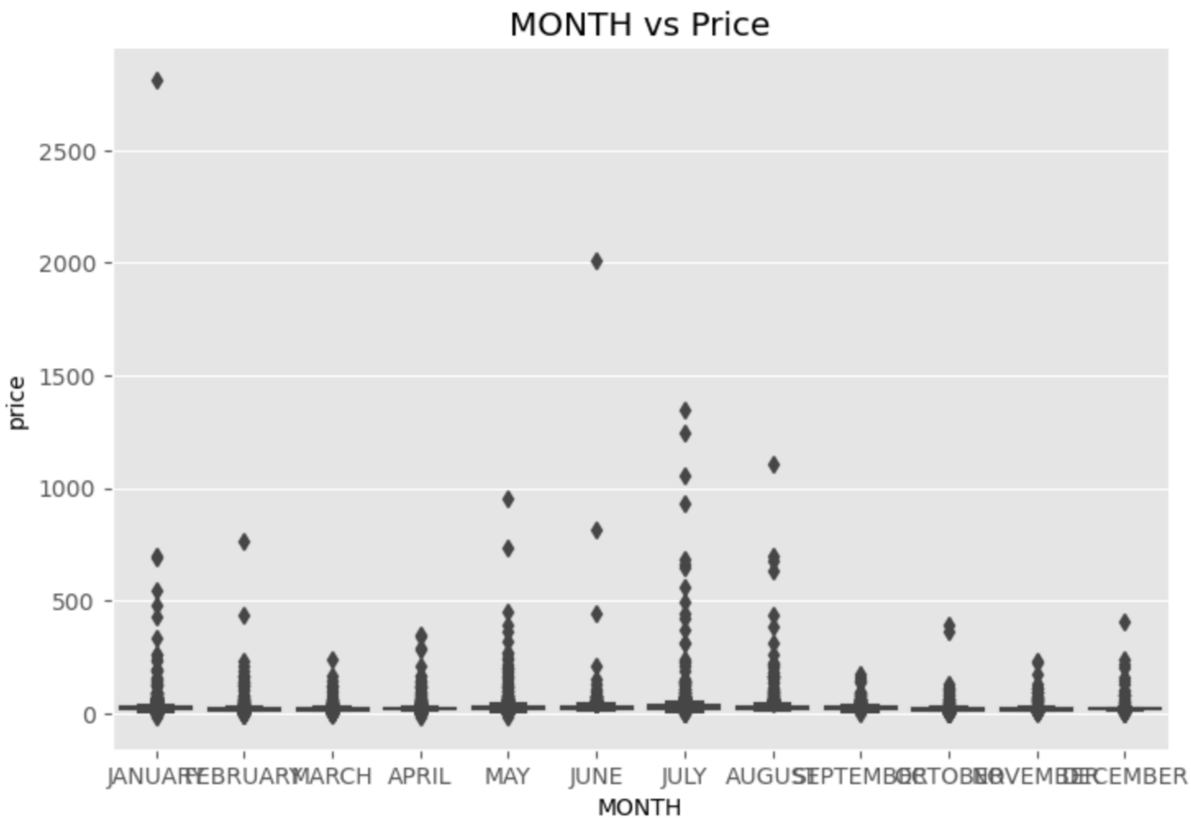
First, we note the object is to predict the price of RTLMP, and we have eight independent variables, wind, solar, and load are numerical variables, and hours ending, peak type, month, and year are categorical variables. Market day is a date variable, so it is tricky; it can be seen as both numerical and categorical variables.

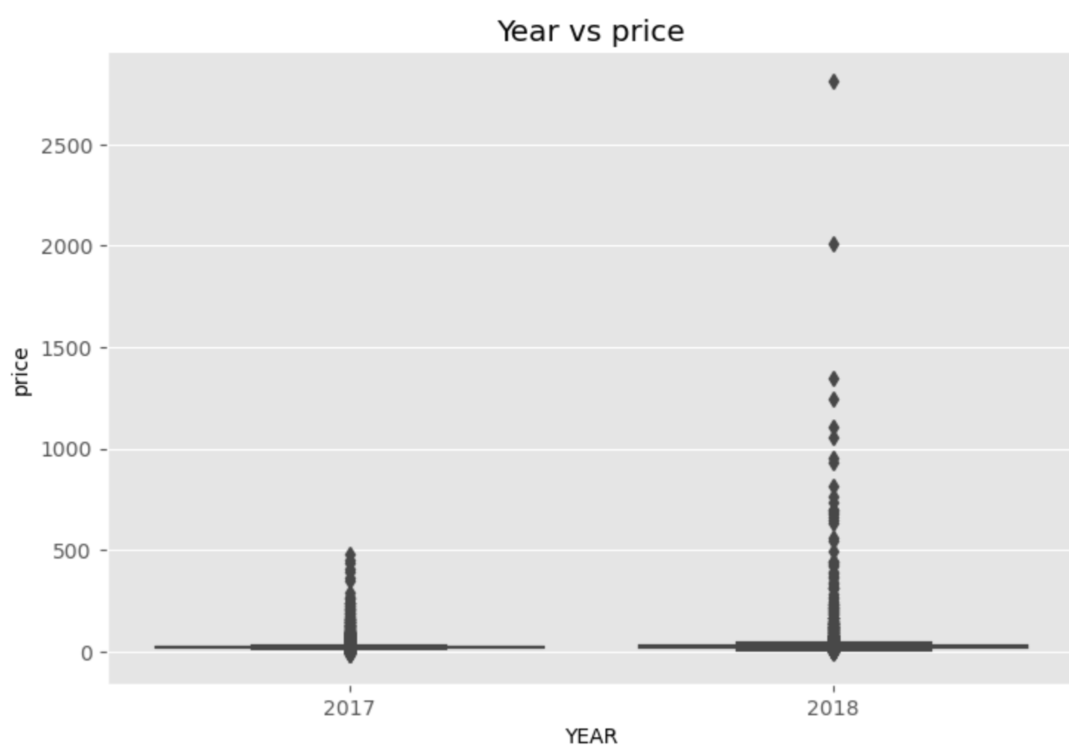
To understand the relationship better, I run a few plot for categorical variables and pair plot for the numerical variables:



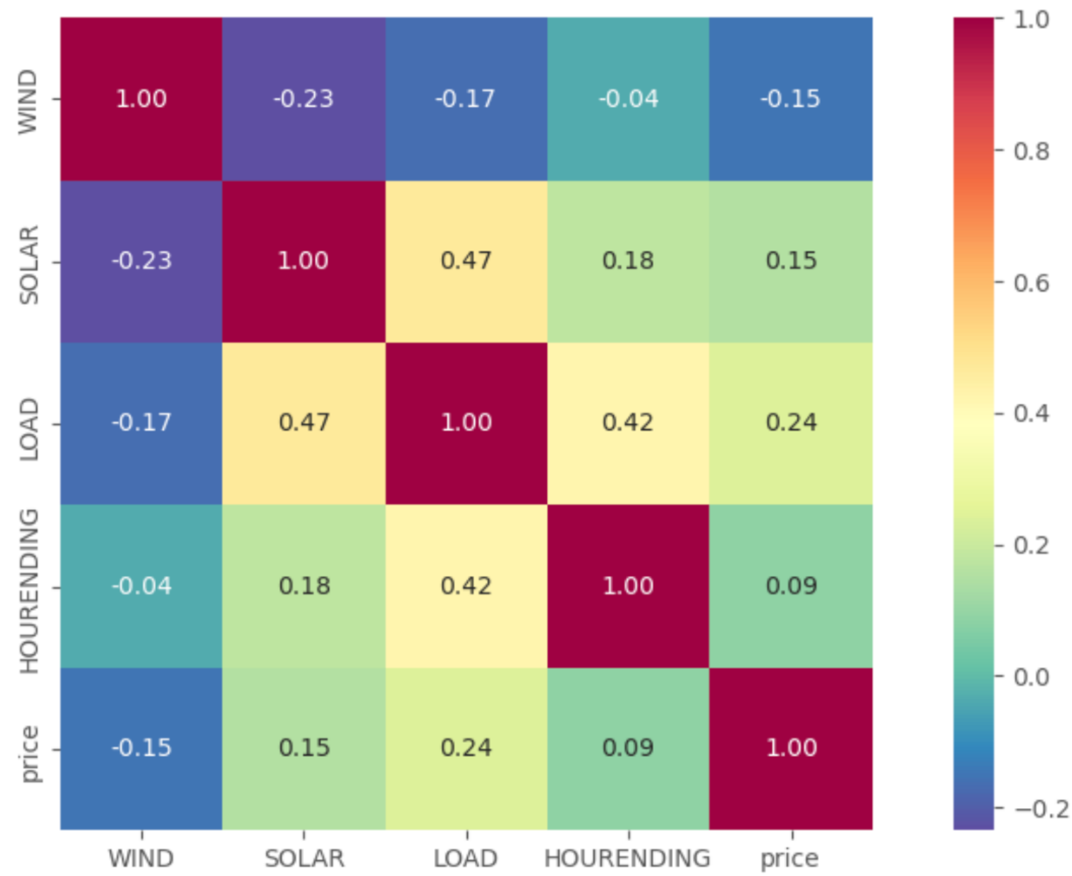

Market day vs Price
price
2500
2000
1500
1000
500
0
MARKETDAY


Peak type vs Price
price
2500
2000
1500
1000
500
0
OFFPEAK
WEPEAK
WDPEAK
PEAKTYPE


MONTH vs Price
price
0
500
1000
1500
2000
2500
JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
MONTH


Year vs price
price
0
500
1000
1500
2000
2500
2017
2018
YEAR

From the categorical variables, it appears that certain market day, hours, month, and peak type has higher prices than others, but these are yet to determine. From the numerical variables, It appears that wind has a negative correlation with the price, while all the other numerical variable has a positive correlation with the price. Furthermore, as the correlation matrix below suggests, there is a high correlation between each independent variable as well, so we might need to remove the collinearity as we are doing the modeling.



The other thing we need to do is to change categorical variables to numerical values, these can be done with some simple data processing knowledge and by creating dummy variables. After removing the collinearity from the independent variables (specifically, we removed to load and market day), we can run a linear regression model to predict the price, given the data:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.050
Model:                            OLS   Adj. R-squared:                  0.049
Method:                 Least Squares   F-statistic:                     46.52
Date:                Mon, 29 May 2023   Prob (F-statistic):         6.79e-153
Time:                        18:11:23   Log-Likelihood:                -78377.
No. Observations:               14987   AIC:                         1.568e+05
Df Residuals:                   14969   BIC:                         1.569e+05
Df Model:                          17
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              29.6934      1.661     17.879      0.000      26.438      32.949
WIND               -0.0015      0.000    -14.415      0.000      -0.002      -0.001
SOLAR               0.0106      0.001      8.323      0.000       0.008       0.013
HOURENDING          0.4010      0.059      6.836      0.000       0.286       0.516
YEAR                5.5073      0.826      6.668      0.000       3.888       7.126
PEAKTYPE_WDPEAK     2.2430      1.082      2.073      0.038       0.122       4.363
PEAKTYPE_WEPEAK     1.6298      1.244      1.310      0.190      -0.808       4.068
MONTH_APRIL        -6.2221      1.688     -3.686      0.000      -9.531      -2.913
MONTH_AUGUST       -3.3763      1.683     -2.006      0.045      -6.675      -0.078
MONTH_DECEMBER     -6.6147      2.072     -3.192      0.001     -10.677      -2.552
MONTH_FEBRUARY     -7.4775      1.702     -4.394      0.000     -10.813      -4.142
MONTH_JULY          0.5551      1.687      0.329      0.742      -2.753       3.863
MONTH_JUNE         -3.2526      1.693     -1.922      0.055      -6.570       0.065
MONTH_MARCH        -9.2147      1.665     -5.534      0.000     -12.479      -5.951
MONTH_MAY          -1.8836      1.675     -1.125      0.261      -5.167       1.399
MONTH_NOVEMBER     -6.4111      2.095     -3.060      0.002     -10.518      -2.305
MONTH_OCTOBER      -6.0652      2.082     -2.913      0.004     -10.147      -1.984
MONTH_SEPTEMBER    -8.8869      1.816     -4.895      0.000     -12.446      -5.328
==============================================================================
Omnibus:                    38905.026   Durbin-Watson:                   1.054
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       1156315640.049
Skew:                          29.875   Prob(JB):                         0.00
Kurtosis:                    1362.463   Cond. No.                     1.00e+05
==============================================================================
```

It looks like we have some insignificant variables with high p-values, so we remove the highest p-value variable each time and run the model again. Finally, we obtain:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.050
Model:                            OLS   Adj. R-squared:                  0.049
Method:                 Least Squares   F-statistic:                     60.32
Date:                Mon, 29 May 2023   Prob (F-statistic):          5.41e-155
Time:                        18:11:36   Log-Likelihood:                -78380.
No. Observations:               14987   AIC:                         1.568e+05
Df Residuals:                   14973   BIC:                         1.569e+05
Df Model:                          13
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              30.0866      1.285     23.423      0.000      27.569      32.604
WIND               -0.0015      0.000    -15.197      0.000      -0.002      -0.001
SOLAR               0.0118      0.001     11.015      0.000       0.010       0.014
HOURENDING          0.4457      0.054      8.210      0.000       0.339       0.552
YEAR                5.4033      0.823      6.569      0.000       3.791       7.016
MONTH_APRIL        -5.7149      1.382     -4.134      0.000      -8.424      -3.005
MONTH_AUGUST       -3.0662      1.361     -2.253      0.024      -5.734      -0.398
MONTH_DECEMBER     -6.1328      1.837     -3.338      0.001      -9.734      -2.531
MONTH_FEBRUARY     -6.8663      1.411     -4.865      0.000      -9.633      -4.100
MONTH_JUNE         -2.8624      1.376     -2.080      0.038      -5.560      -0.165
MONTH_MARCH        -8.6235      1.362     -6.331      0.000     -11.293      -5.954
MONTH_NOVEMBER     -5.9239      1.859     -3.186      0.001      -9.568      -2.279
MONTH_OCTOBER      -5.6374      1.838     -3.067      0.002      -9.241      -2.034
MONTH_SEPTEMBER    -8.5656      1.531     -5.596      0.000     -11.566      -5.565
==============================================================================
Omnibus:                    38906.519   Durbin-Watson:                   1.055
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       1157816648.568
Skew:                          29.877   Prob(JB):                         0.00
Kurtosis:                    1363.347   Cond. No.                     6.07e+04
==============================================================================
```

Now this model has all variables being significant, from the model result, we can conclude that:
- Wind has a negative influence on the price, specifically -0.0015 per unit of increase in wind
- Solar has a positive influence on the price, specifically 0.0118 per unit of increase in solar
- Later ending hour has a positive influence on the price, with a coefficient of 0.4457 per hour
- The year 2018 has a higher price than the year 2017, on average by 5.4033

On average, compared to the month of **January**:

- The price in April is 5.71 lower
- The price in August is 3.07 lower
- The price in December us 6.13 lower
- The price in February is 6.87 lower
- The Price in June is 2.86 lower
- The Price in March is 8.62 lower
- The price in November is 5.92 lower
- The price in October is 5.64 lower
- The price in September is 8.57 lower
- There is no significant difference in price between January and May or July

- Also, there is no significant price difference in regards to peak type (WEPEAK, WDPEAK and OFFPEAK)