Daily electricity Price and Demand Analysis

Statistical Learning Final Project

Mojtaba Amini
Javier Alberto Bernal Sigala
Carmen Rocio Ortiz Benitez
Saeed Soufeh

Outline

- Objectives
- Data Collection
- Exploratory Data Analysis (EDA)
- Model Data & Analysis
- Model Evaluation
- Conclusion

Objectives

- Explore the data in depth in order to understand relations between variables and their behavior.
- Build a regression model to predict electricity demand.
- Build a logistic regression model to classify the RRP as above or below the median.

Data Collection

- Daily electricity price and demand in the state of Victoria, Australia
- 2,106 observations , 14 variables
- Duration: January 2015 to December 2020

Dataset Variables(1/2)

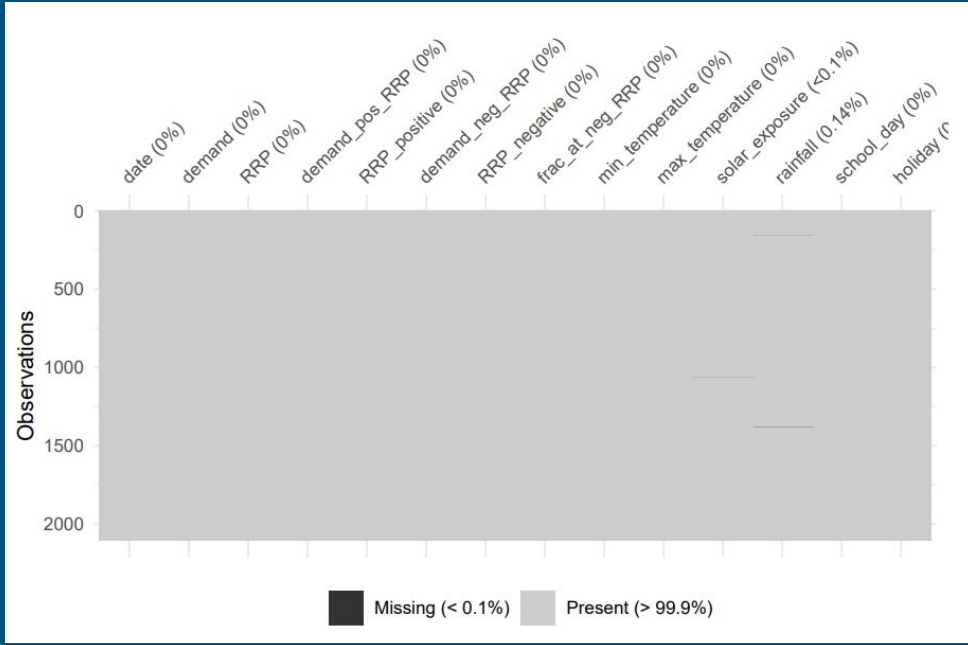
- demand: total daily electricity demand in MWh (megawatt hour);
- RRP: (a recommended retail price in AUD$ / MWh (Australian Dollars per MWh);
- demand_pos_RRP: total daily demand at positive RRP in MWh;
- RRP_positive: averaged positive RRP, weighted by the corresponding daily demand in AUD$ / MWh
- demand_neg_RRP: total daily demand at negative RRP in MWh (megawatt hour);
- RRP_negative: average negative RRP, weighted by the corresponding daily demand in AUD$ / MWh
- frac_at_neg_RRP: fraction of the day when the demand was traded at negative RRP;

Dataset Variables(2/2)

- min_temperature: minimum temperature during the day in Celsius;
- max_temperature: maximum temperature during the day in Celsius;
- solar_exposure: total daily sunlight energy in MJ/mˆ2 (megajoule per square meter);
- rainfall: rainfall during the day in mm;
- school_day: whether students were at school on that day;
- holiday: whether the day was a holiday.

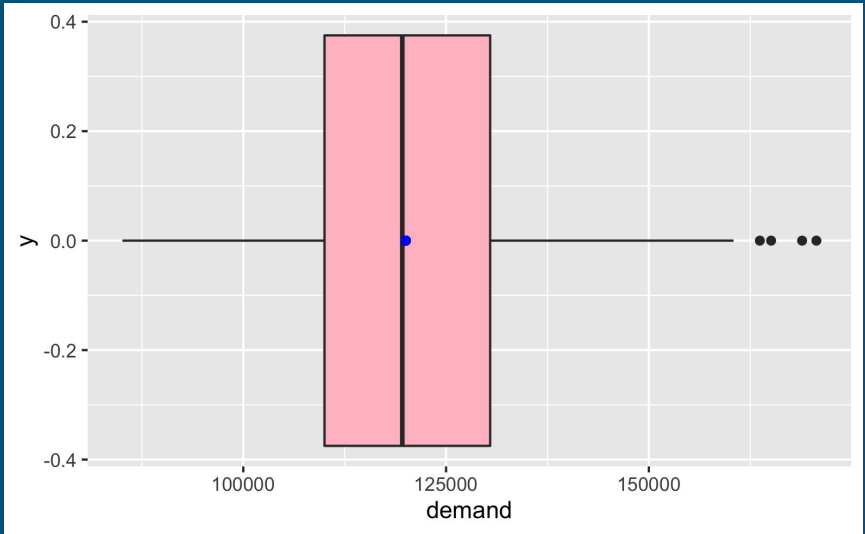
Data Preparation

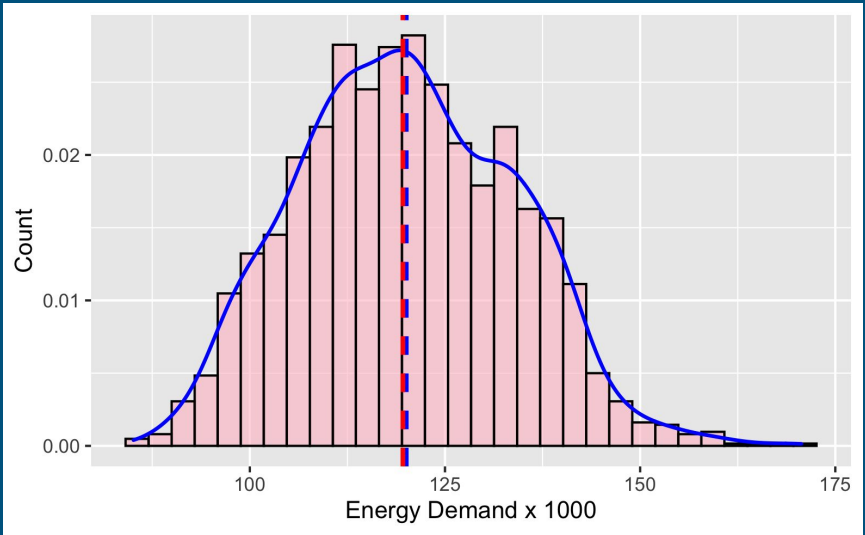
- Check for duplicated rows (there were not any duplicated rows)
- Check for missing values and remove them
- transform the date column from character to date
- the min and max temperature into one combined mean temperature
- Change holiday and school_day columns from character to categorical (binary)



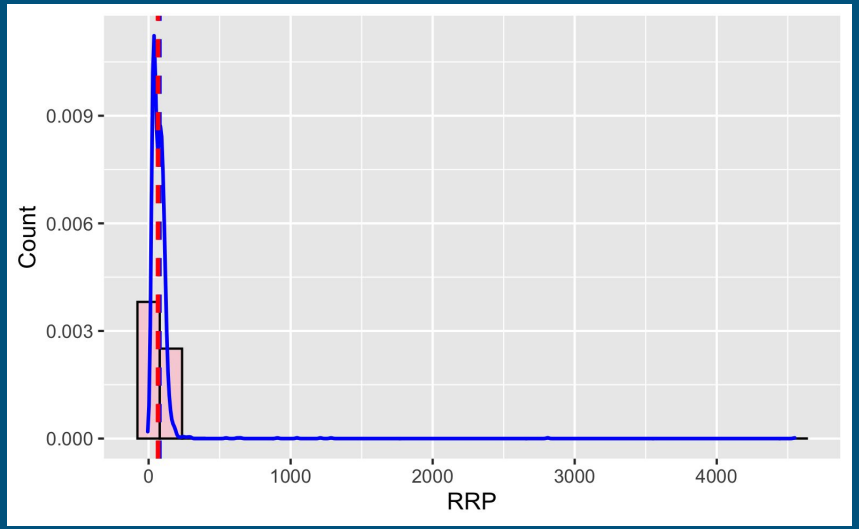
Exploratory Data Analysis (EDA)

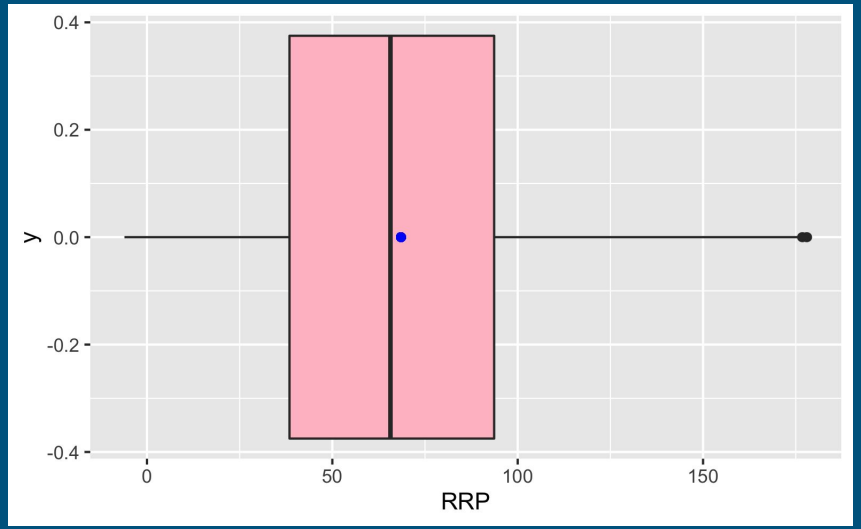
Demand (total daily electricity demand in MWh (megawatt hour))





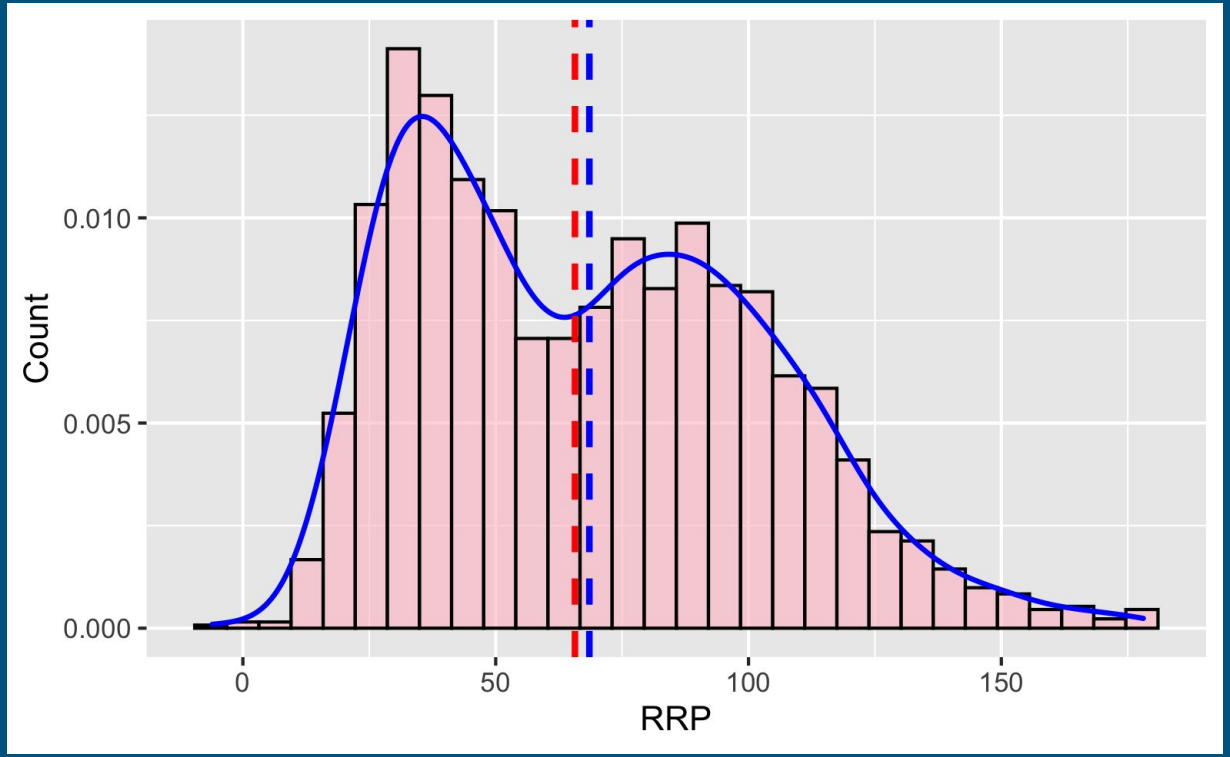
RRP (a recommended retail price in AUD$ / MWh (Australian Dollars per MWh)





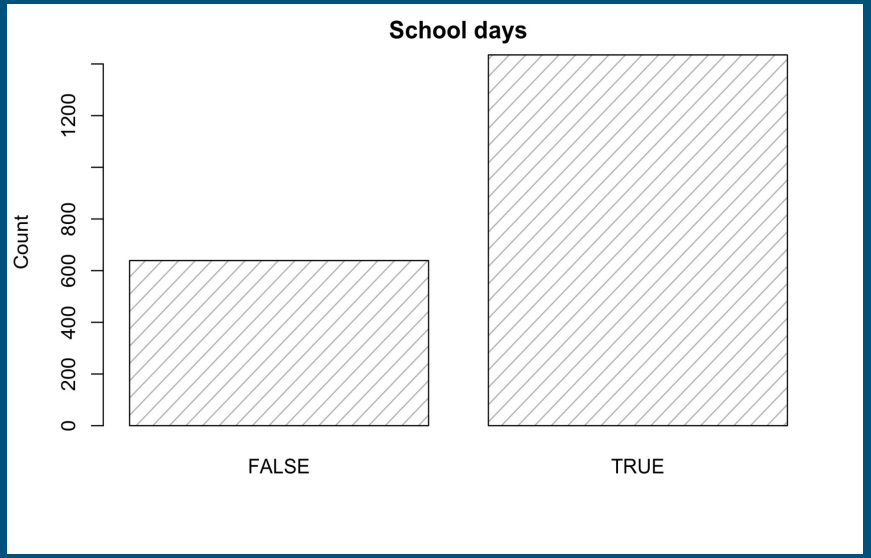
RRP

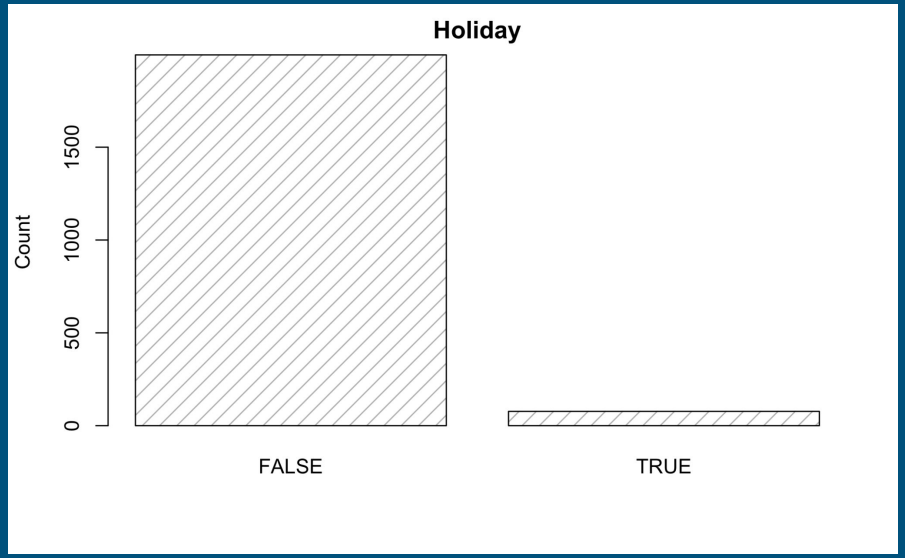
Without the outliers, RRP is shown to have a bimodal distribution,



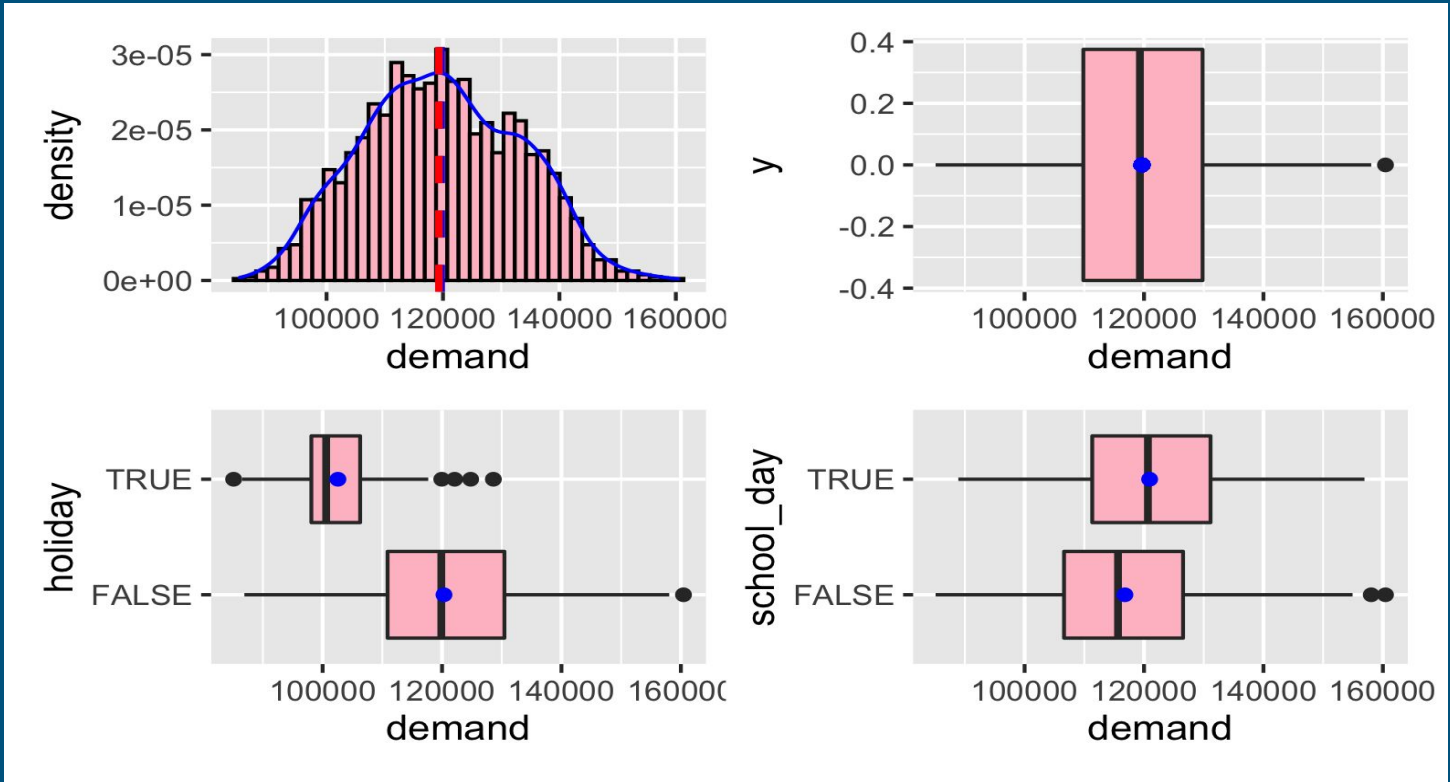
School Days (whether students were at school on that day)

Holidays (whether the day was a holiday)



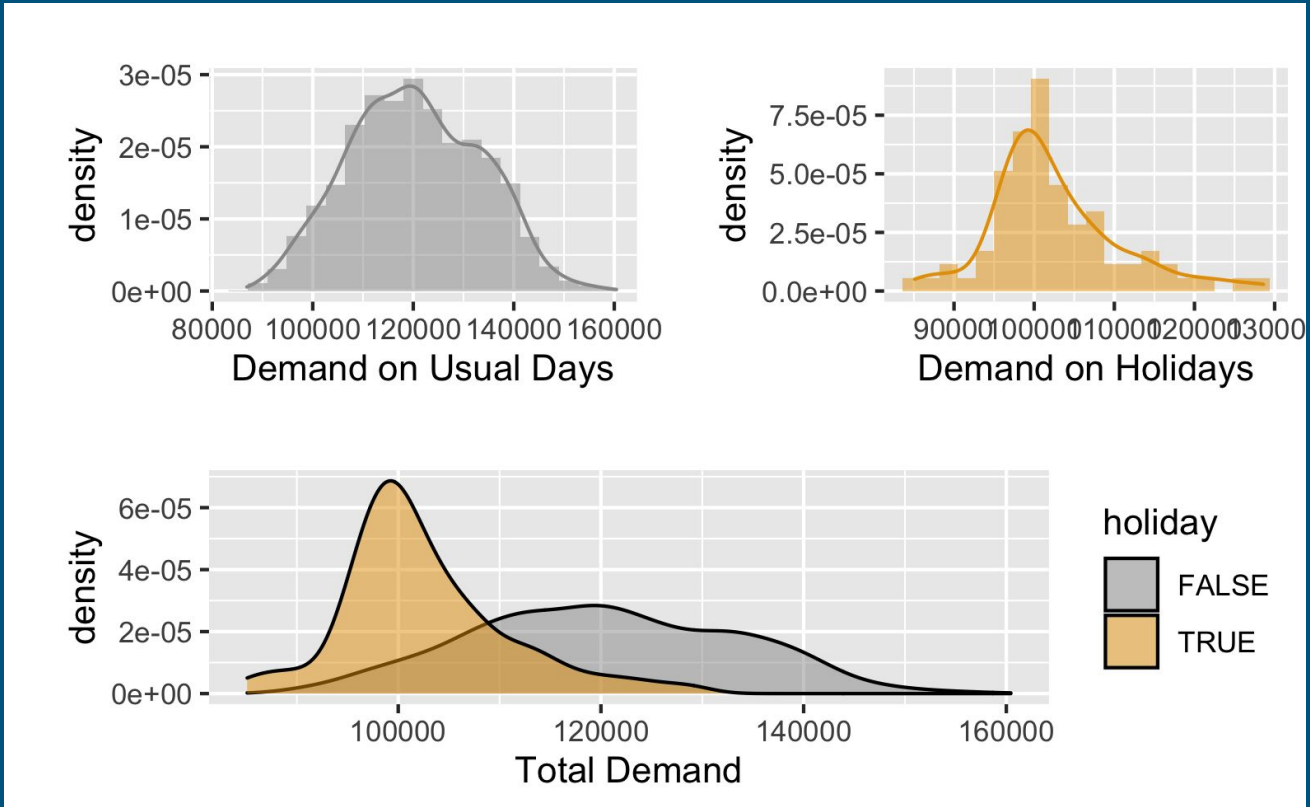


The distribution of electricity demand on its own is almost symmetrical.



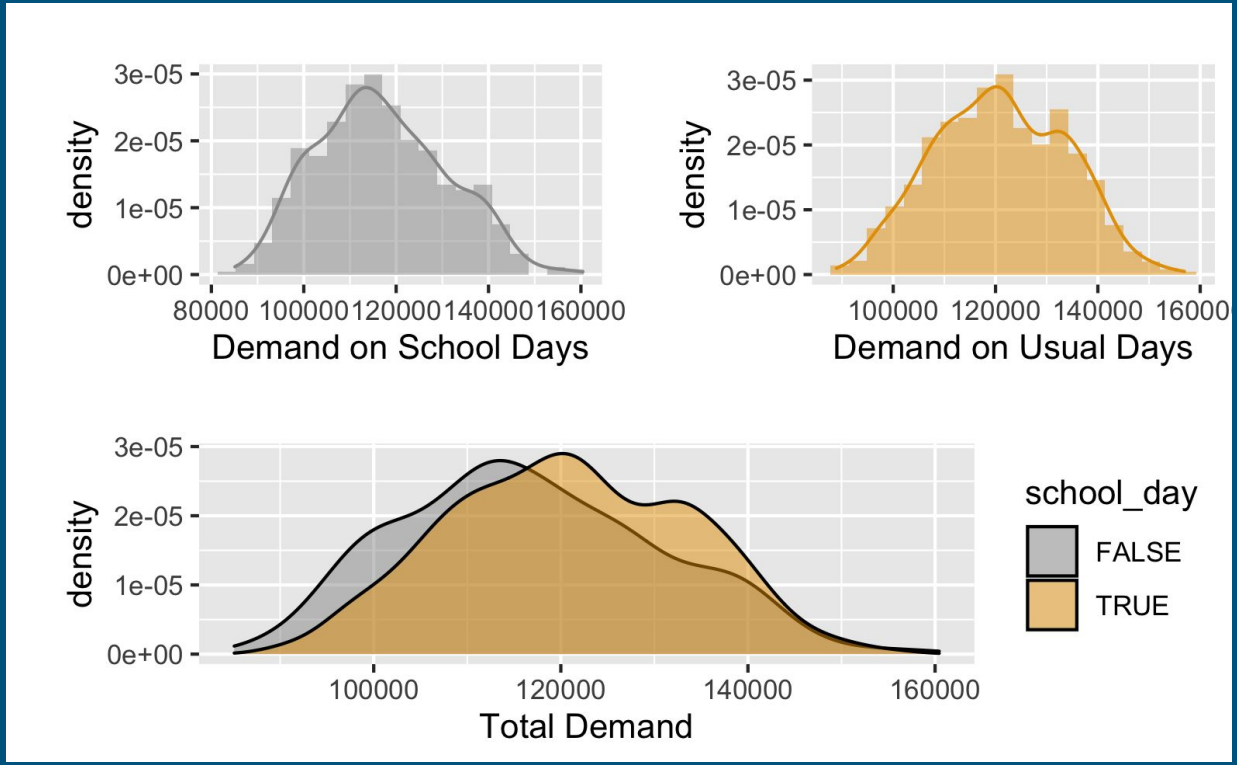
Demand & Holidays

The difference in distributions in demand for holidays and non-holidays is evident.



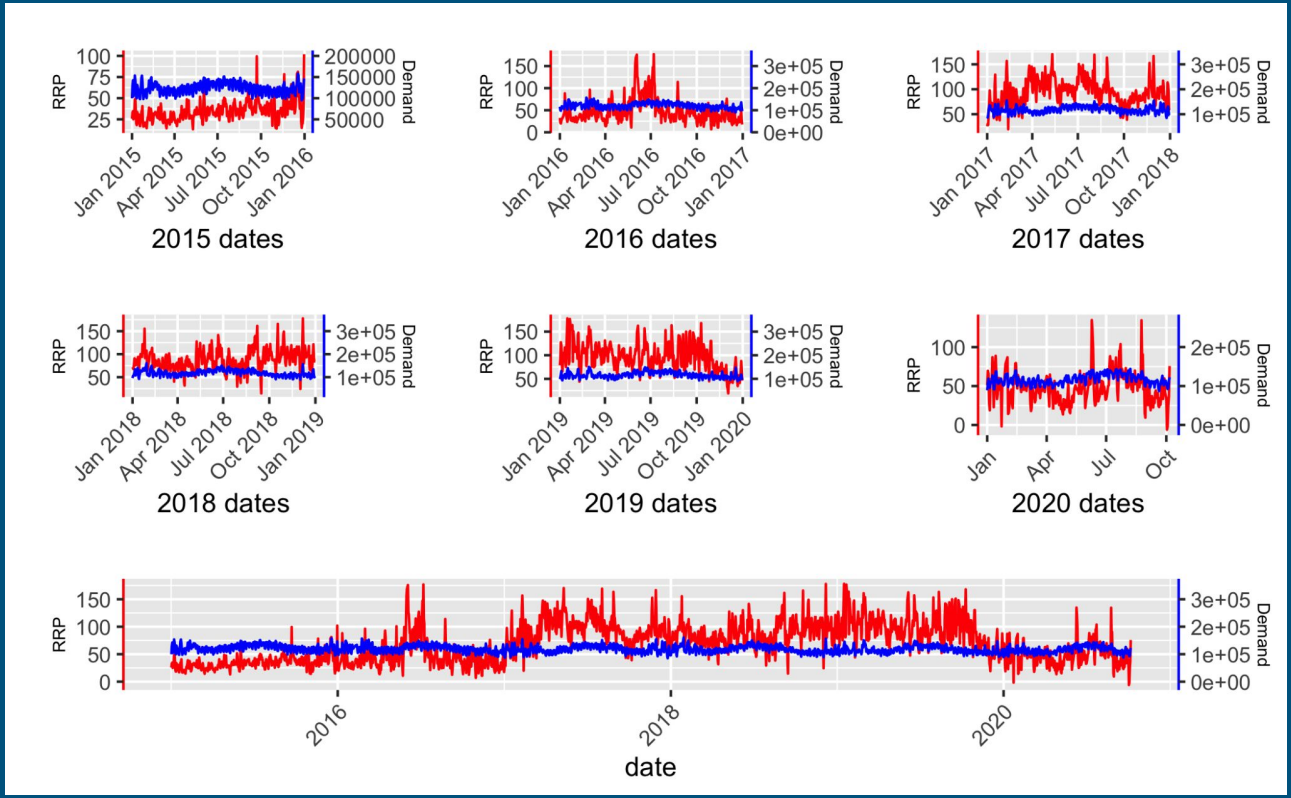
Demand & School days

Mean demand on school days is a little higher, but the variance does not differ by a lot.



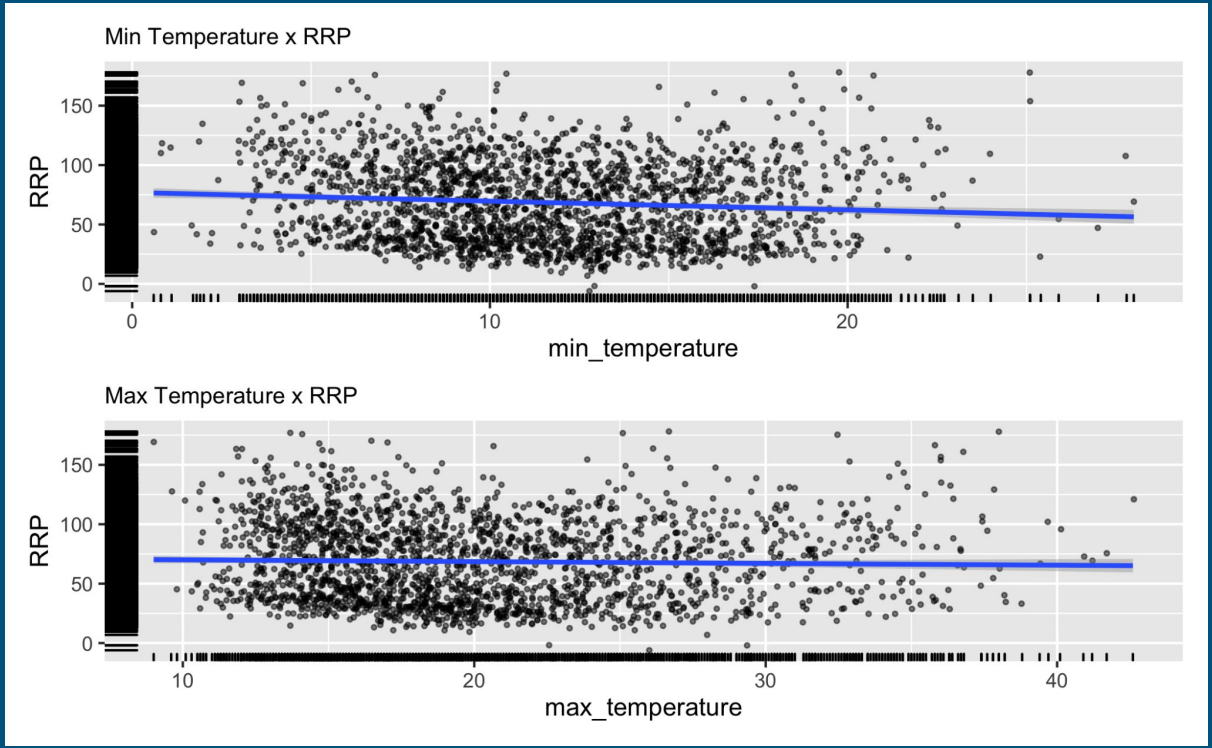
Date

While demand does not fluctuate much over the years, RRP does shift considerably during and throughout the years.



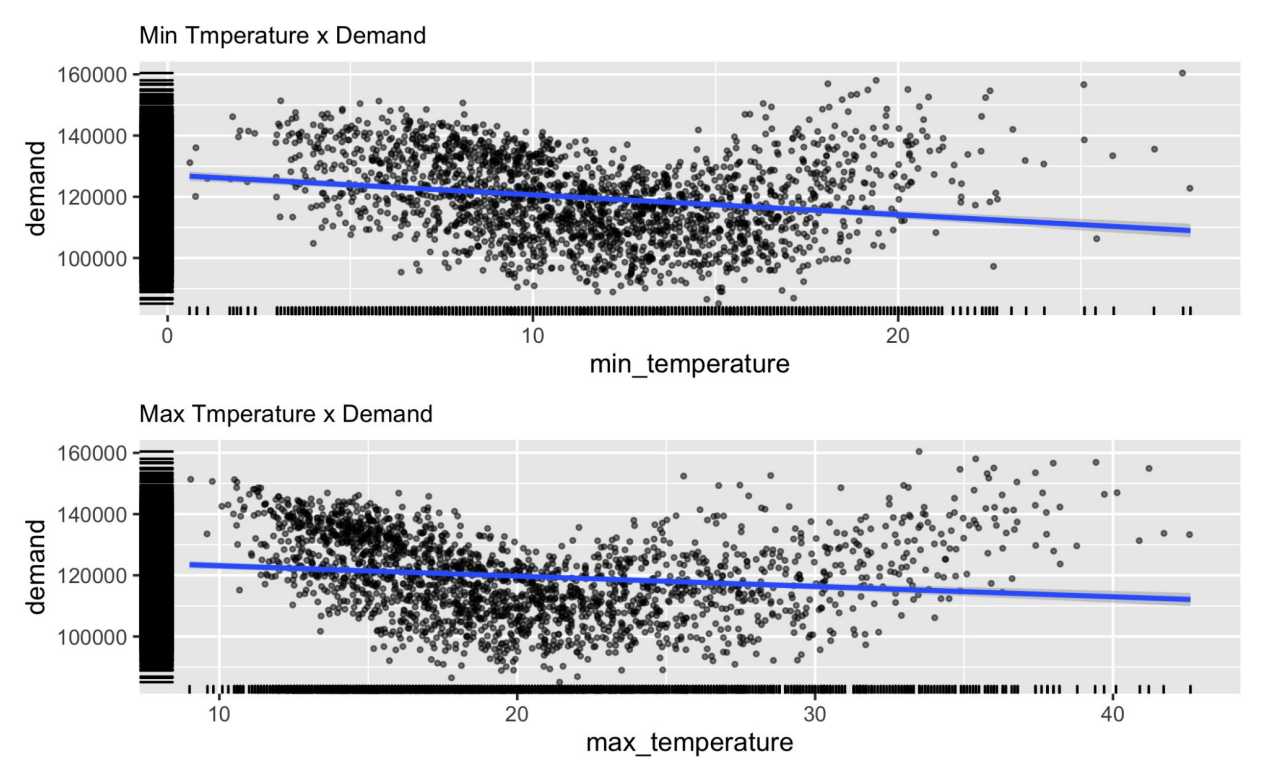
Temperatures & RRP

The relations of min and max temperature with RRP are extremely similar.



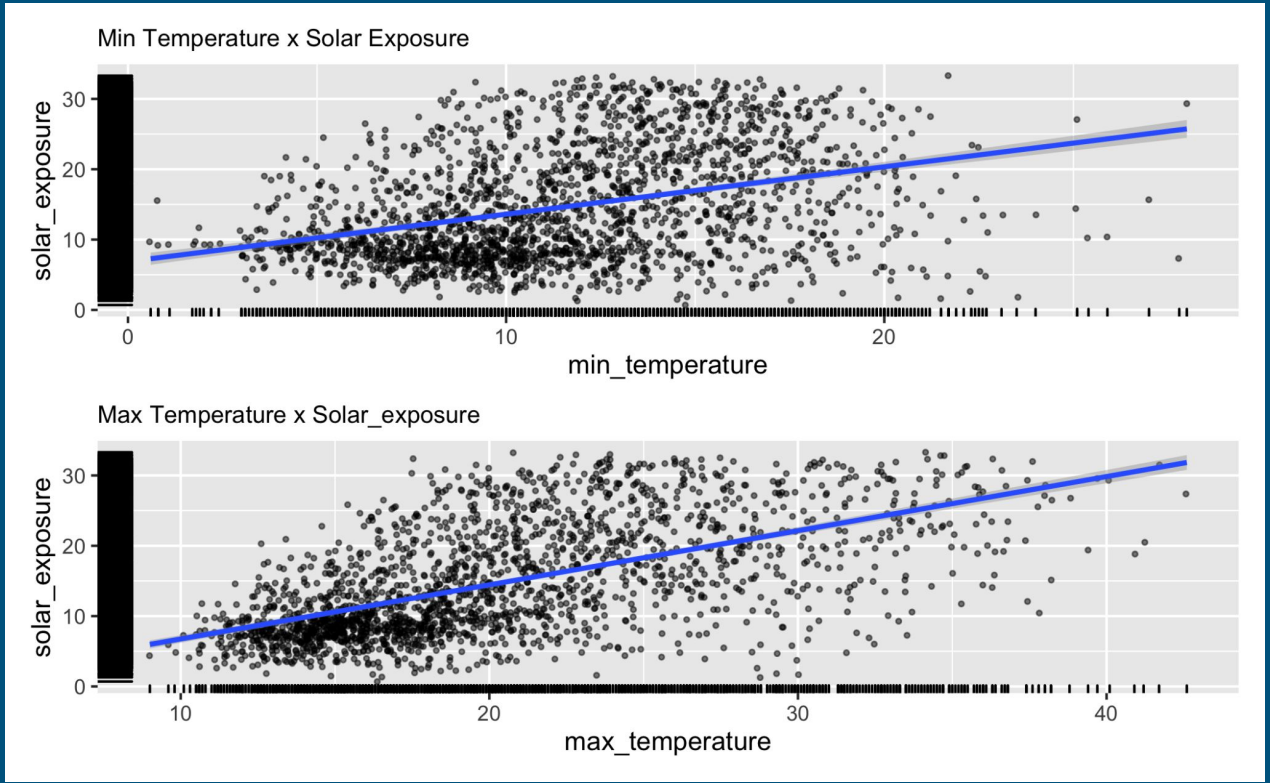
Temperatures & Demand

The similarity in relations here is a bit less than with RRP, but still existent.



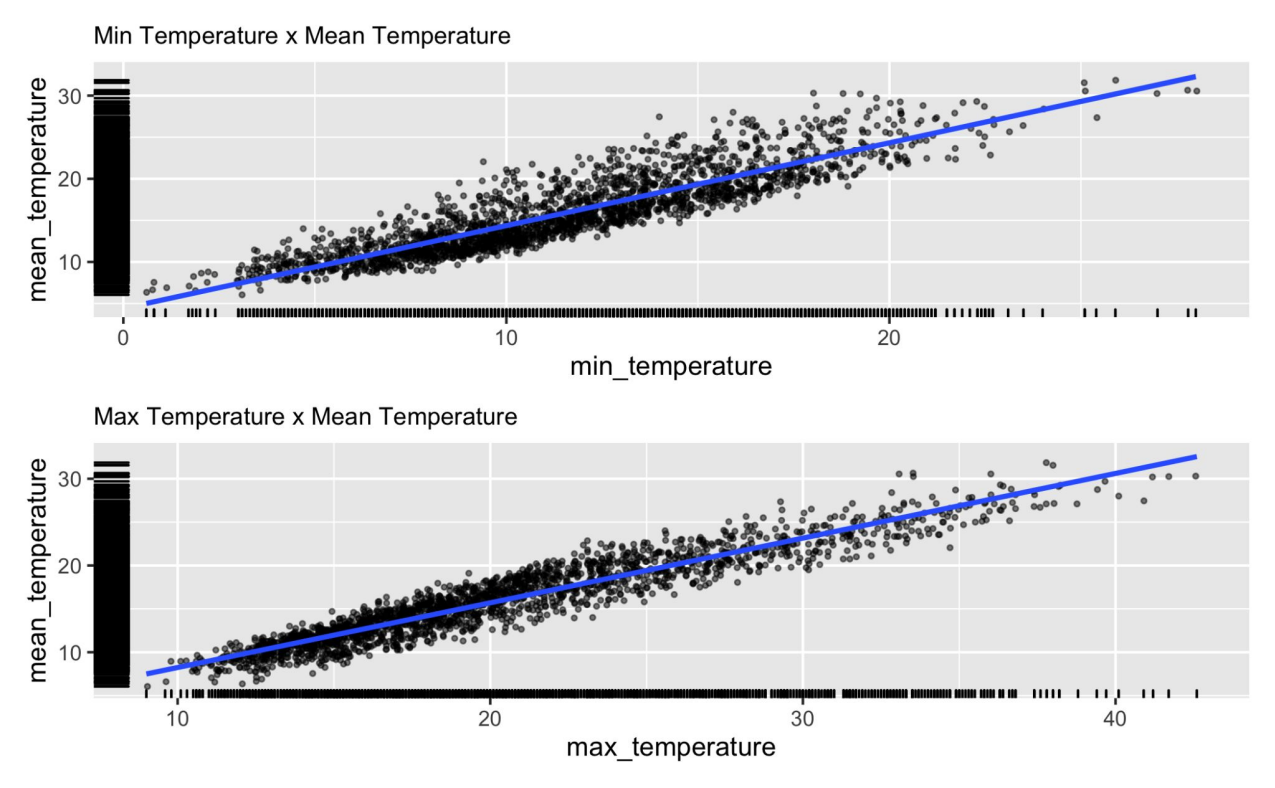
Temperatures & Solar Exposure

Here, they are extremely similar again.



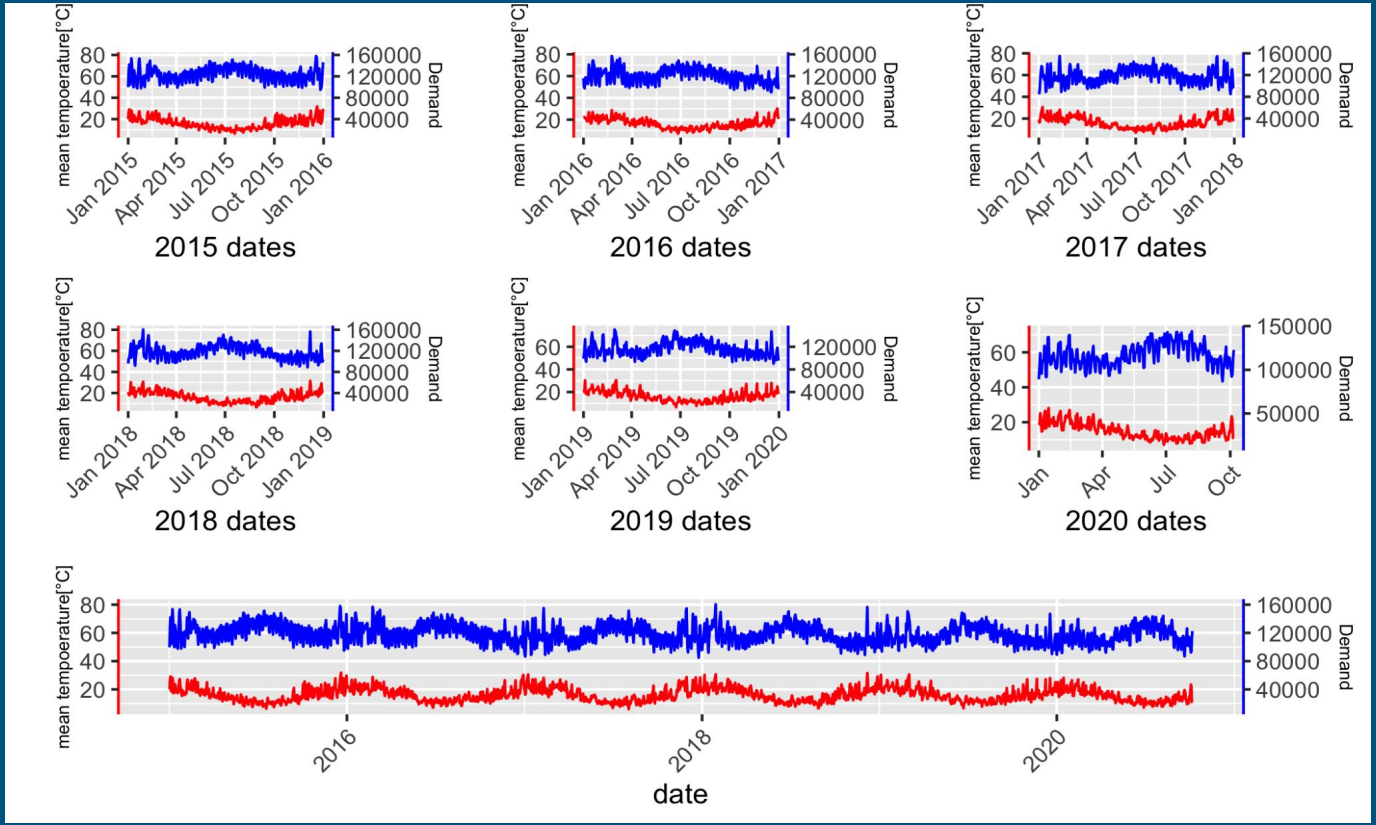
Temperatures & Mean Temperature

Again, very similar relations. We conclude that the min and max temperature can be combined into one single mean temperature variable.



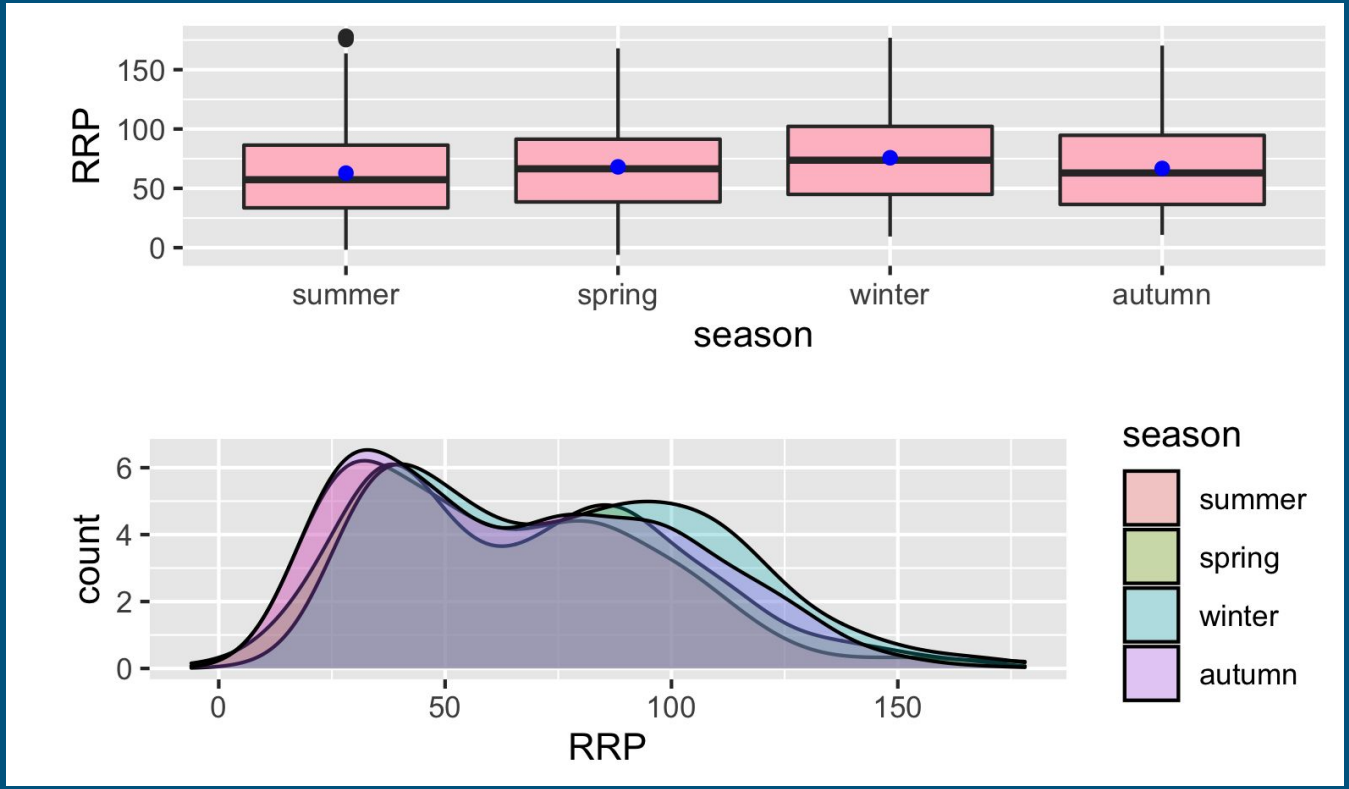
Demand & Mean Temperature by dates

When mean temperature peaks, demand goes down and vice versa.



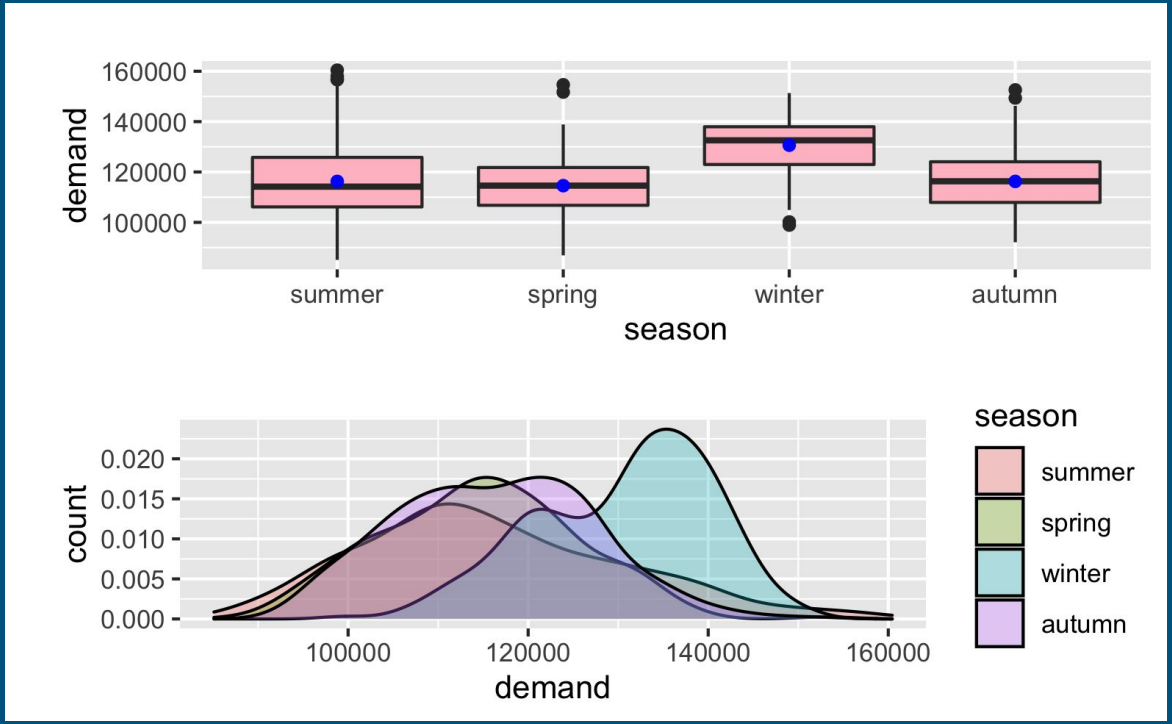
Seasons & RRP

RRP values are the highest in winter and the lowest in summer.



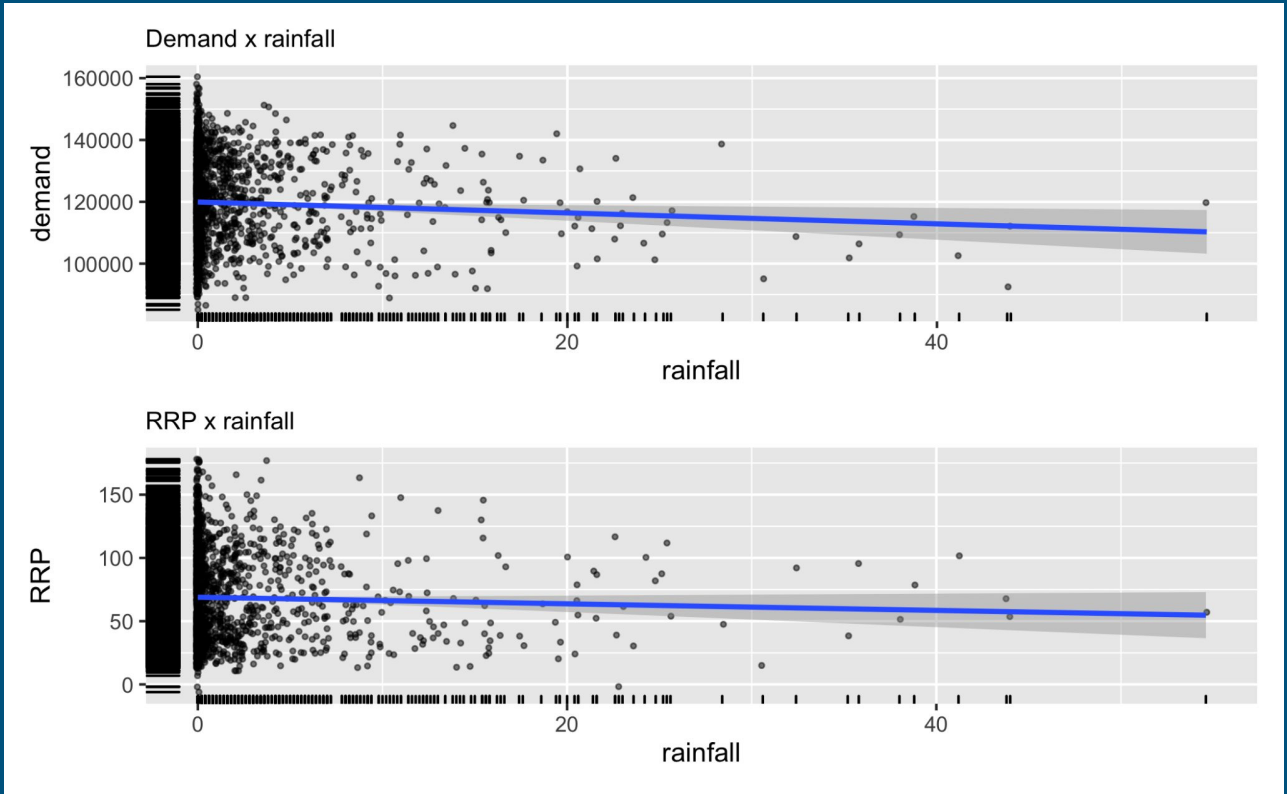
Seasons & Demand

Demand on winter is higher than in the other seasons. Summer has the lowest demand comparatively.



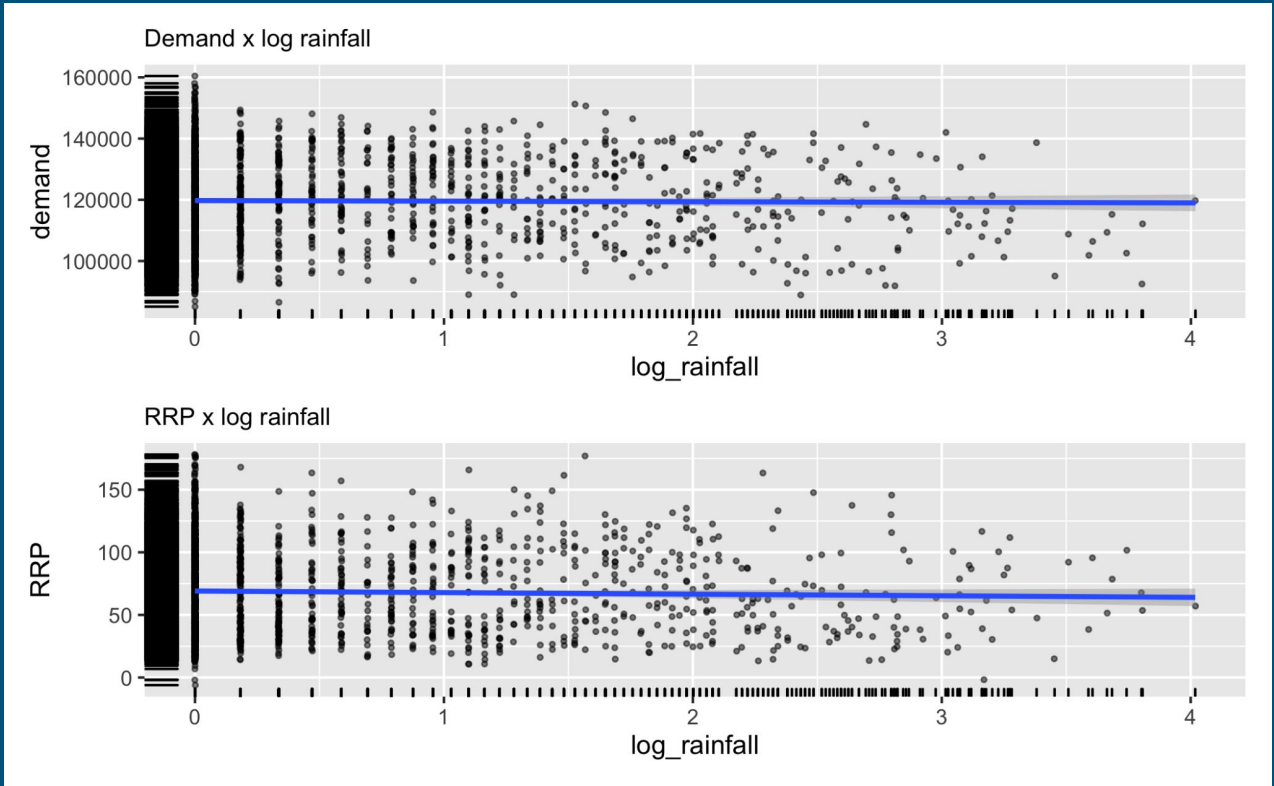
Demand & Rainfall + RRP & Rainfall

As rainfall increases, demand and RRP decrease, although not much.



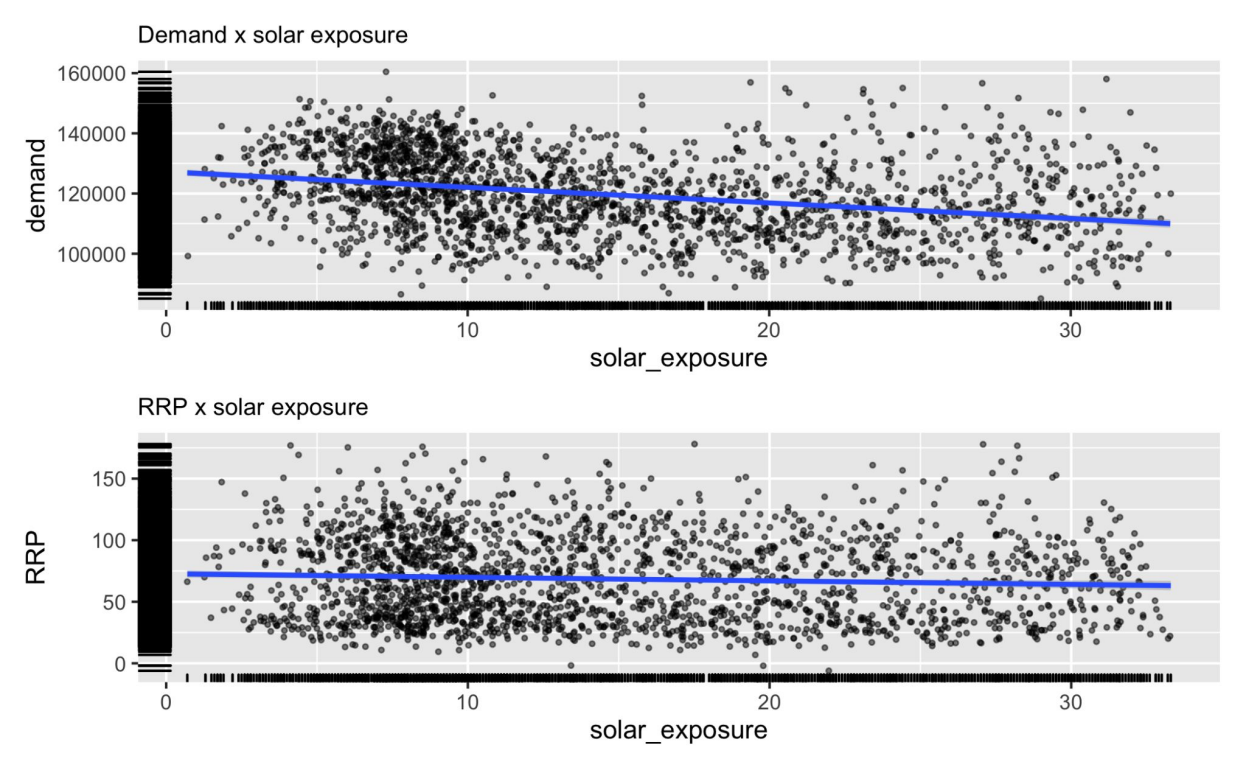
Demand & Log Rainfall + RRP & Log Rainfall

After the transformation of rainfall, the effect in demand and RRP is still not significant, especially for demand. RRP decreases very slightly as log rainfall increases.



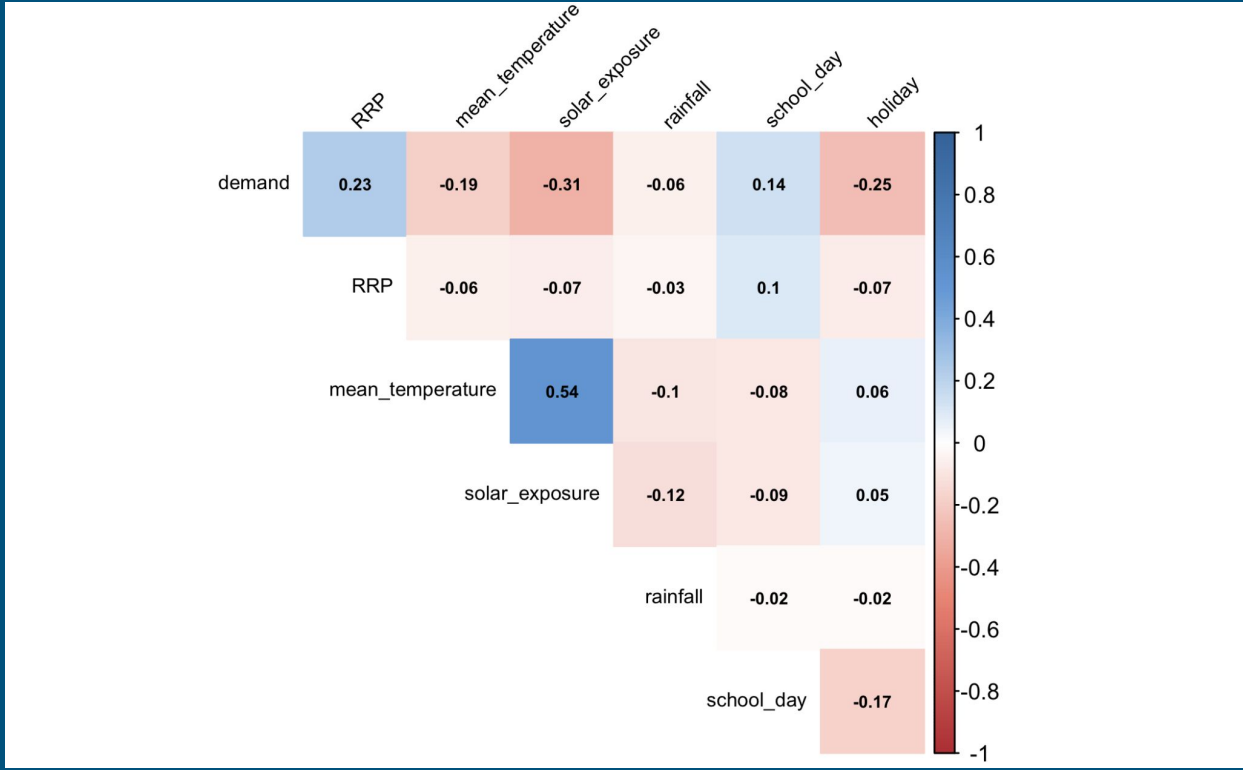
Demand & Solar Exposure + RRP & Solar Exposure

As seen below, as solar exposure increases demand decreases. As for RRP, the effect of solar exposure is almost imperceptible



Correlations

Most important correlations are between demand and solar_exposure, and demand and holiday.The feature that has less correlation with demand is rainfall.



Variable statistics

Hypothesis 1

- Goal: T-test for the equality of means of the demand on school days and non-school days
- Prerequisite: the equality of the variances (we can examine this through an F-test)

- H_0: $\sigma_{school} \neq \sigma_{notschool}$
- H_1: $\sigma_{school} \neq \sigma_{notschool}$

We cannot perform a t-test to analyze the equality of the mean, since we do not have evidence to conclude that the variances of demand on school day vs not school days are equal.

```
##  F test to compare two variances
##
## data:  school_days and no_school_days
## F = 0.84801, num df = 1434, denom df = 638, p-value = 0.01305
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7418882 0.9660391
## sample estimates:
## ratio of variances
##          0.8480142
```

Hypothesis 2

- Goal: T-test for the equality of means of the demand on holidays and non-holidays
- Prerequisite: the equality of the variances (we can examine this through an F-test)

- H_0: $\sigma_{holiday} \neq \sigma_{notholiday}$
- H_1: $\sigma_{holiday} \neq \sigma_{notholiday}$

Again, we cannot conclude that the variances of the two populations are equal and thus we cannot perform a t-test

```
##  F test to compare two variances
##
## data:  holidays and no_holidays
## F = 0.38706, num df = 76, denom df = 1996, p-value = 7.539e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2863989 0.5497053
## sample estimates:
## ratio of variances
##          0.3870597
```

Model Data & Analysis

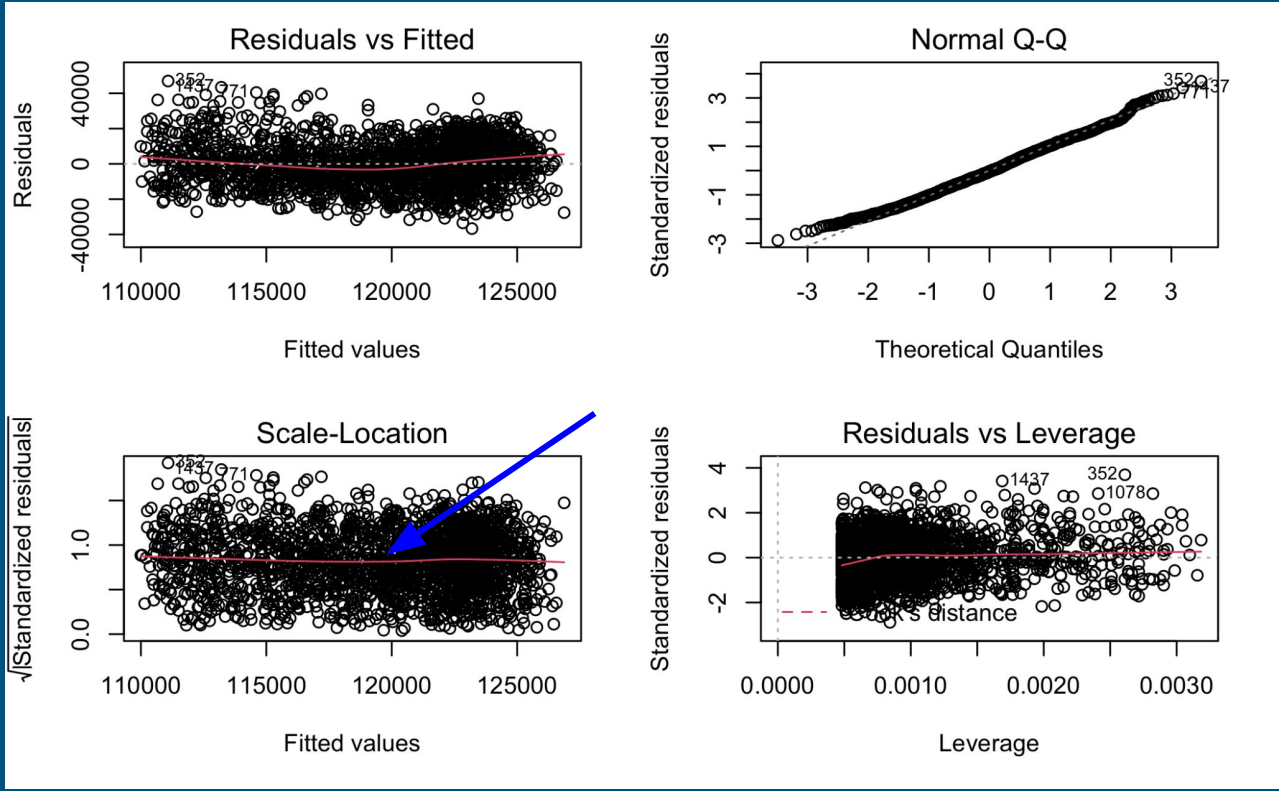
Linear Regression Model:

Numerical variables:

- In this section we will consider the effect of numerical variables on electricity demand.
- As we know from the correlation matrix performed during the EDA, the solar exposure (solar_exposure) has high correlation with demand. So, we start by building a linear regression model with only this variable.
- Performance in this section will be analyzed in terms of BIC ((Bayesian Information Criterion).

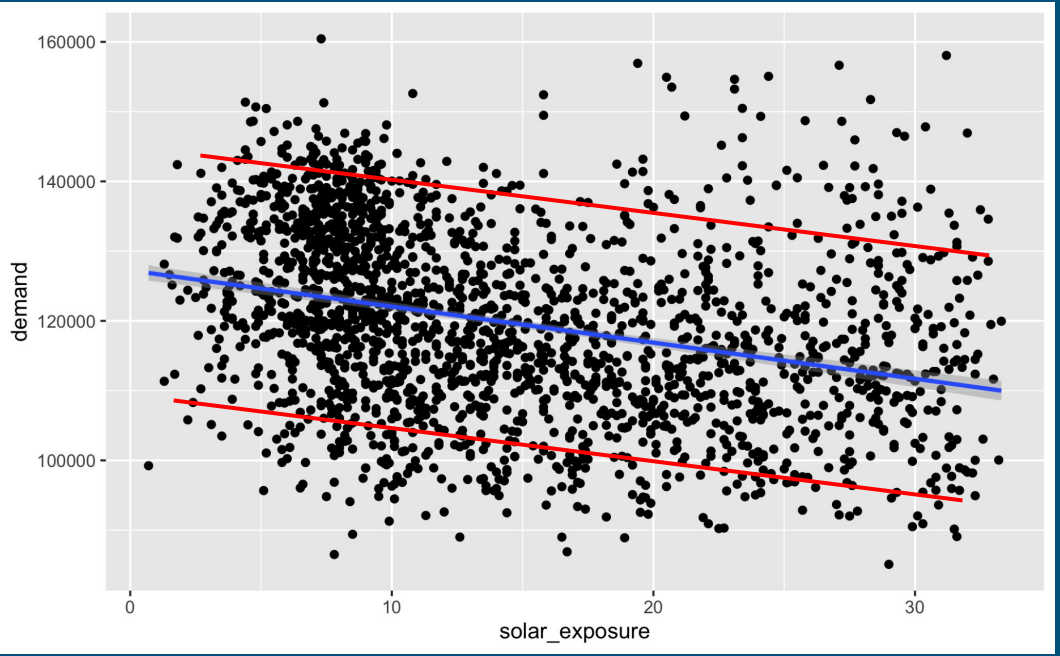
Regression Model 0 - demand ~ solar_exposure

The behavior of the model is acceptable, since residuals are constant and normally distributed



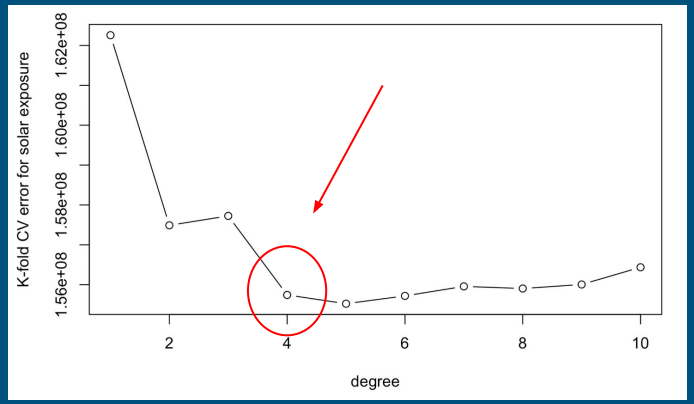
Regression Model 0 - demand ~ solar_exposure

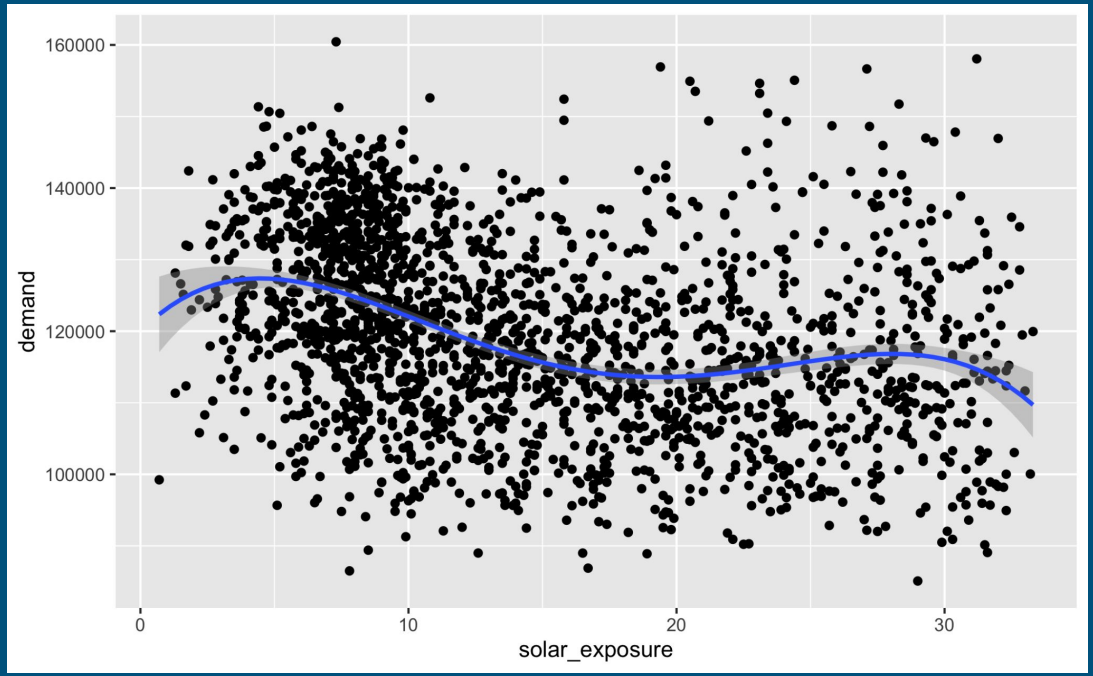
- The shape of this model.
- From the below plot, we can see that our model may benefit by increasing the polynomial degree.



Regression Model 1 - demand ~ poly(solar_exposure,4)

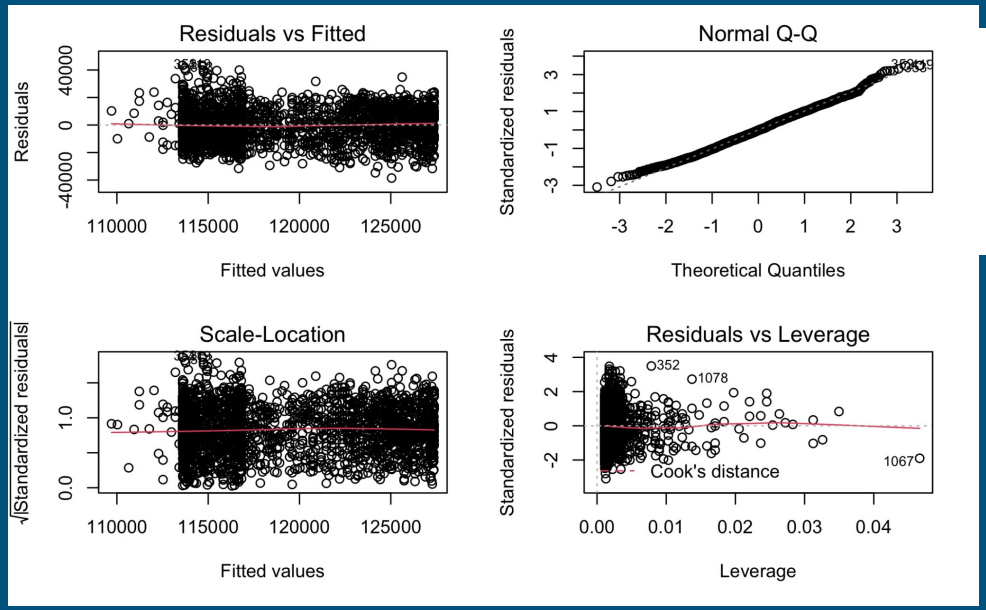
After comparing different degree models, we find that the best performance is obtained by the fourth degree one.



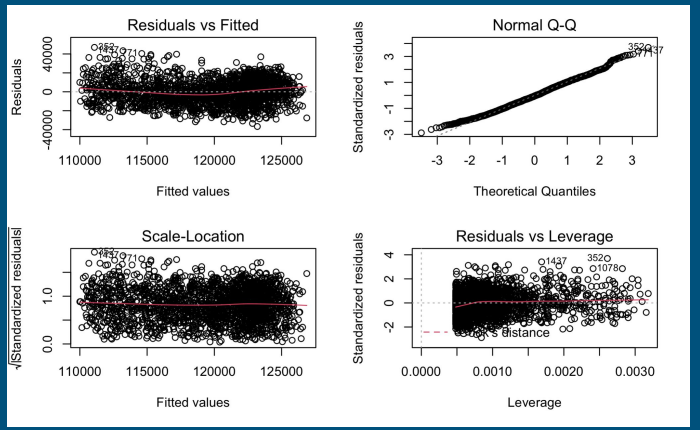


Regression Model 1 - demand ~ poly(solar_exposure, 4)

The behavior of residuals does not change much but BIC is a little worse

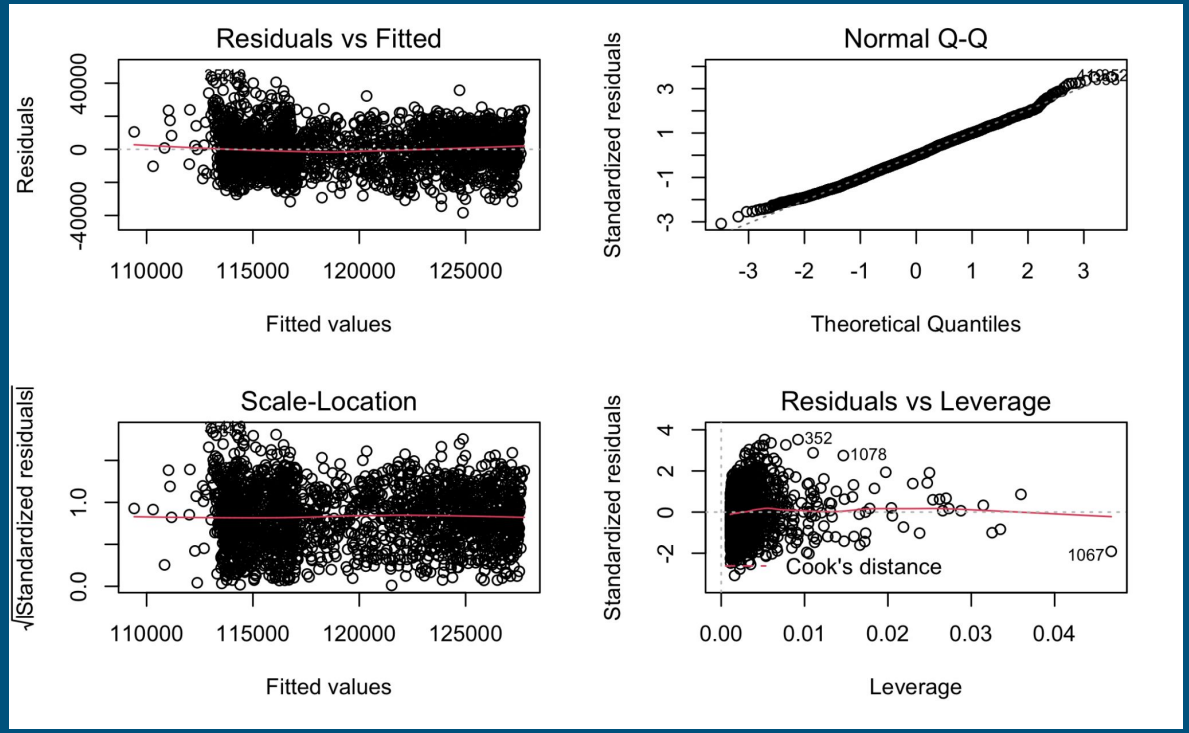


Variables_num	BIC_num	Adjusted_R2_num
solar exposure^4/mean temperature	45119.29	0.1308000
solar exposure^4/mean temperature^4	44083.89	0.4601000
solar exposure^4/mean temperature^4/log rainfall	44066.24	0.4664000
solar exposure^4	45044.64	0.1310000
solar exposure	45113.74	0.0928700
mean temperature^4	44193.91	0.4234000
mean temperature	45241.01	-0.0354500



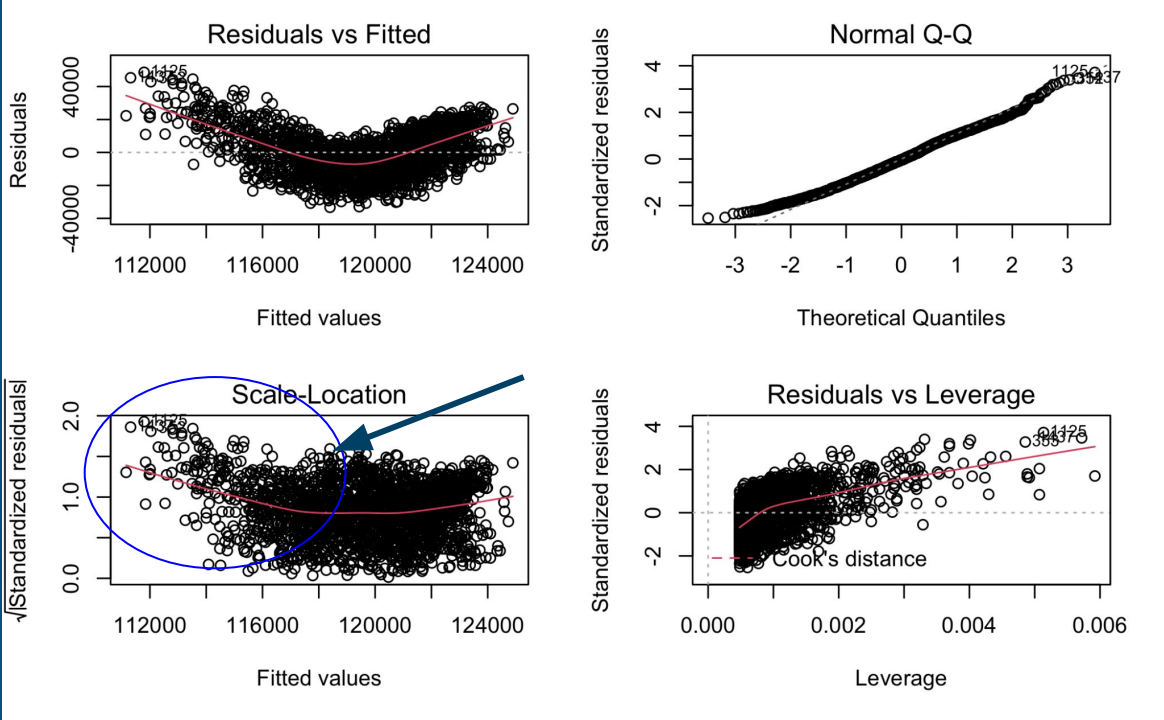
Regression Model 2 - demand ~ poly(solar_exposure, 4) + mean_temperature

For this model the behavior of residuals is still acceptable but BIC is a little worse.



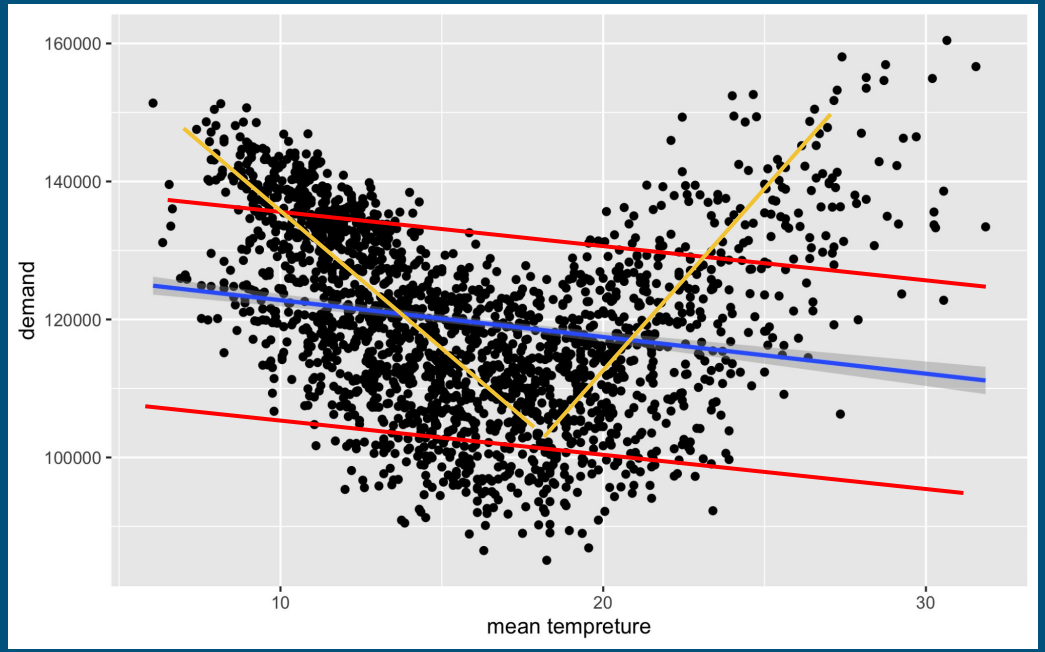
Regression Model 3 - demand ~ mean_temperature

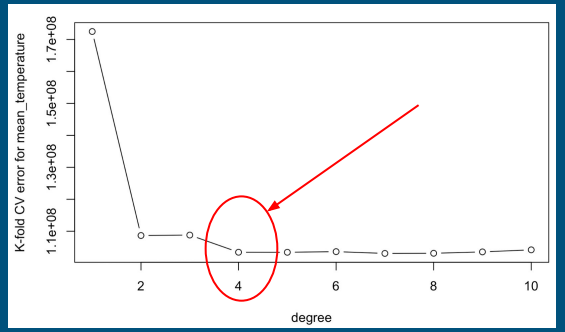
Check the behavior of the model with only the mean temperature variable in order to have a better idea of the effect of this variable

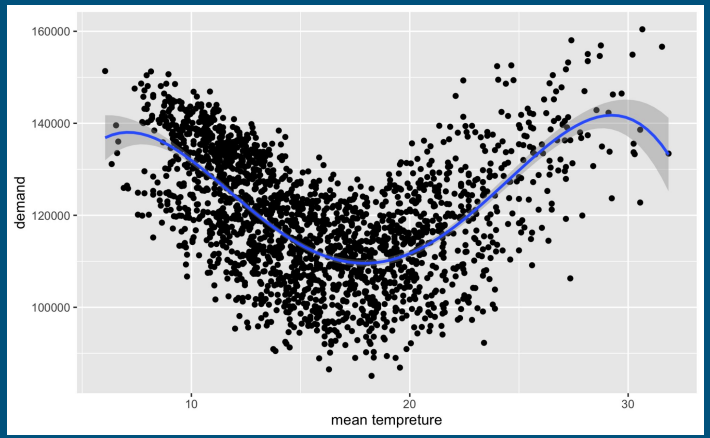


Regression Model 3 - demand ~ mean_temperature

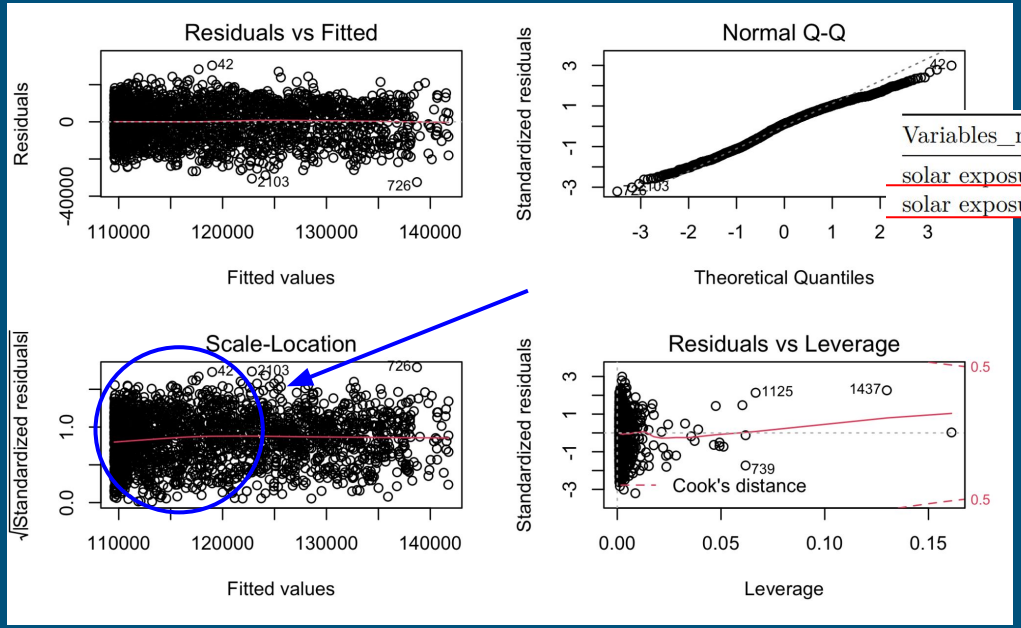
The behavior of mean temperature as seen below suggests that increasing the degree of the mean temperature variable can improve the model



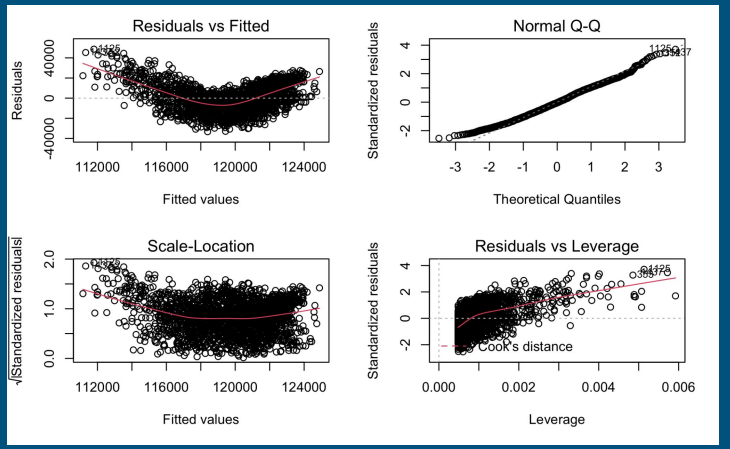




Regression Model 4 - demand ~ poly(mean_temperature, 4)

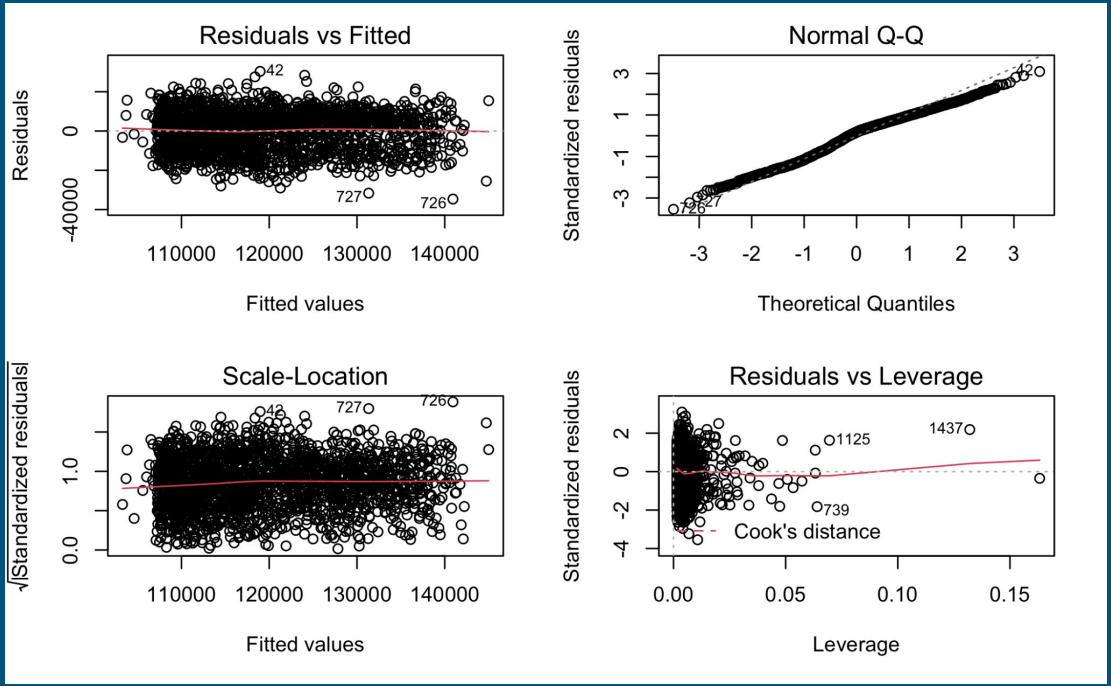


Variables_num	BIC_num	Adjusted_R2_num
solar exposure^4/mean temperature	45119.29	0.1308000
solar exposure^4/mean temperature^4	44083.89	0.4601000



Regression Model 5 - demand ~ poly(solar_exposure, 4) + poly(mean_temperature, 4)

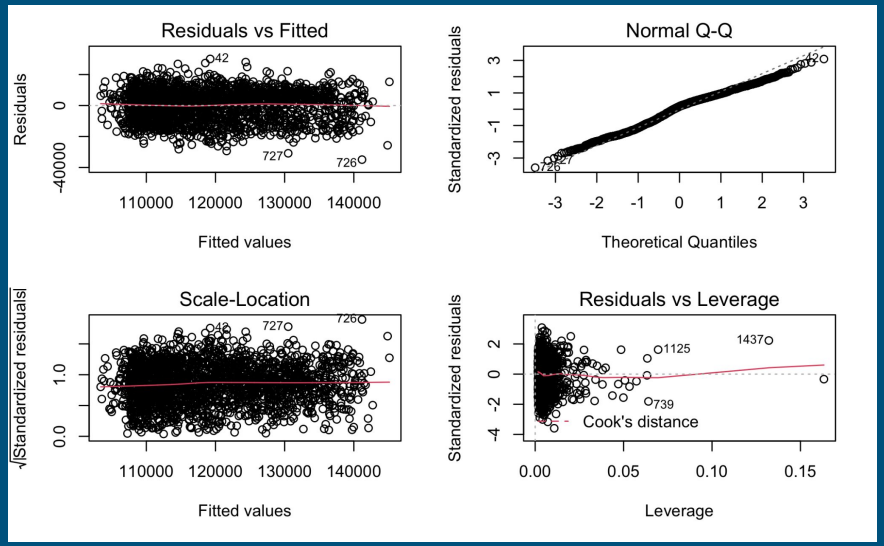
Residual behavior is also acceptable. Since the model behavior improved, we keep this fourth degree term and go on to add a variable that we previously examined on the EDA section, log_rainfall.

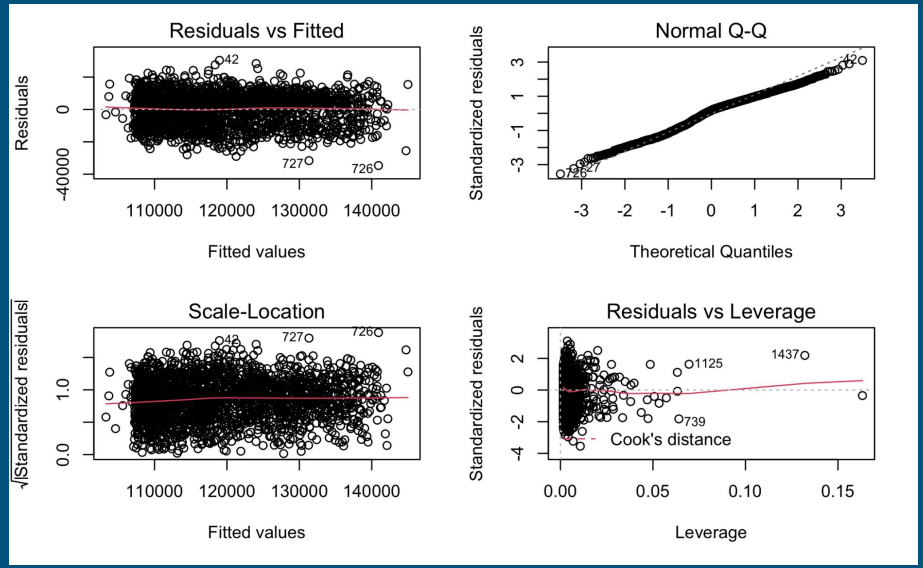


Regression Model 6 - demand ~ poly(solar_exposure, 4) + poly(mean_temperature, 4) + log_rainfall

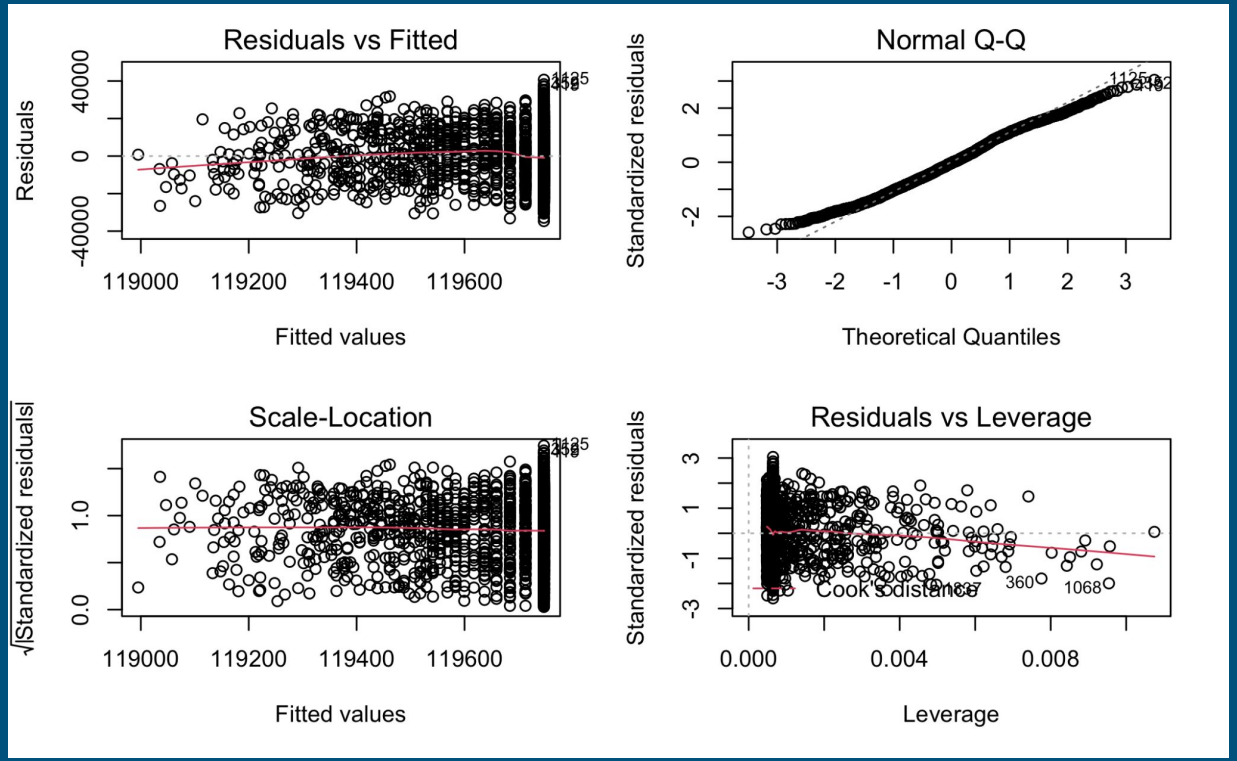
Residual plots show an admissible behavior. Now, we check the behavior of the model with only the log_rainfall variable.

Variables_num	BIC_num	Adjusted_R2_num
solar exposure^4/mean temperature	45119.29	0.1308000
solar exposure^4/mean temperature^4	44083.89	0.4601000
solar exposure^4/mean temperature^4/log rainfall	44066.24	0.4664000





Regression Model 7 - demand ~ log_rainfall



Metrics (conclusion about the numerical variables)

Variables_num	BIC_num	Adjusted_R2_num
solar exposure^4/mean temperature	45119.29	0.1308000
solar exposure^4/mean temperature^4	44083.89	0.4601000
solar exposure^4/mean temperature^4/log rainfall	44066.24	0.4664000
solar exposure^4	45044.64	0.1310000
solar exposure	45113.74	0.0928700
mean temperature^4	44193.91	0.4234000
mean temperature	45241.01	-0.0354500
log rainfall	45316.64	-0.0003652

Categorical variables

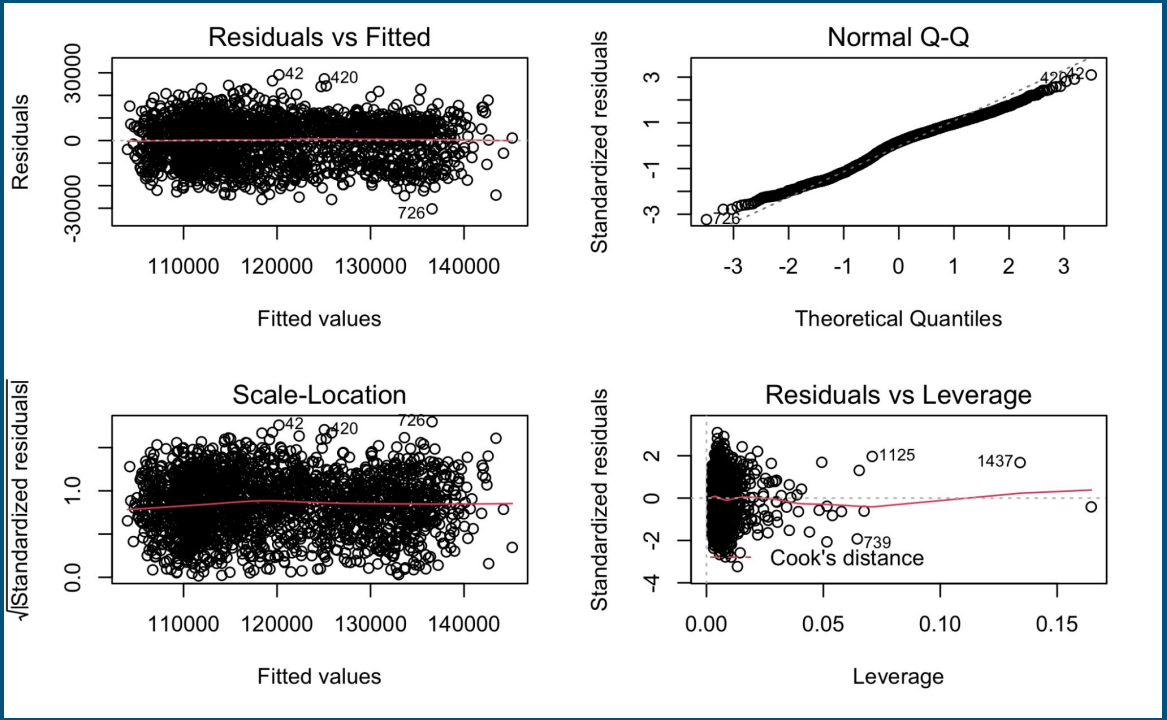
- We try all possible combinations of the variables to find the model which results in the best performance in terms of BIC and Adjusted R^2.
- Since the holiday data is too unbalanced, we remove the feature and keep only the “school_day” and “season”.

variables	BIC	adjusted_R2
school day/season/holiday	44616.54	0.2953041
school day/season	44724.16	0.2553871
season/holiday	44626.09	0.2897857
school day/holiday	45166.55	0.0724428
school day	45274.07	0.0199582
holiday	45181.64	0.0626738
season	44750.67	0.2433993

Unified model

demand ~ poly(solar_exposure, 4) + poly(mean_temperature, 4) + season + school_day

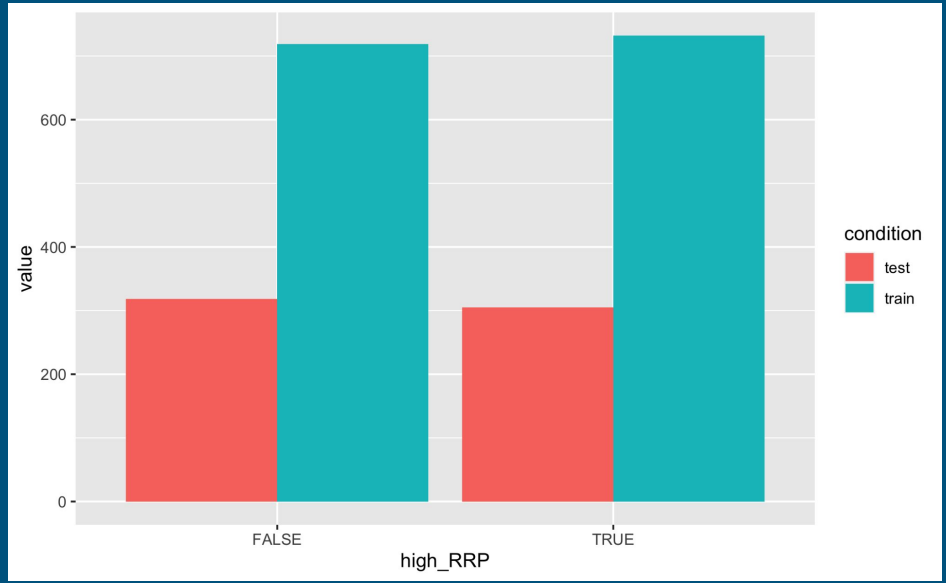
Adjusted R-squared: 0.5029 **BIC: 43939.4**



Model Data & Analysis (second part)

Logistic Regression Model:

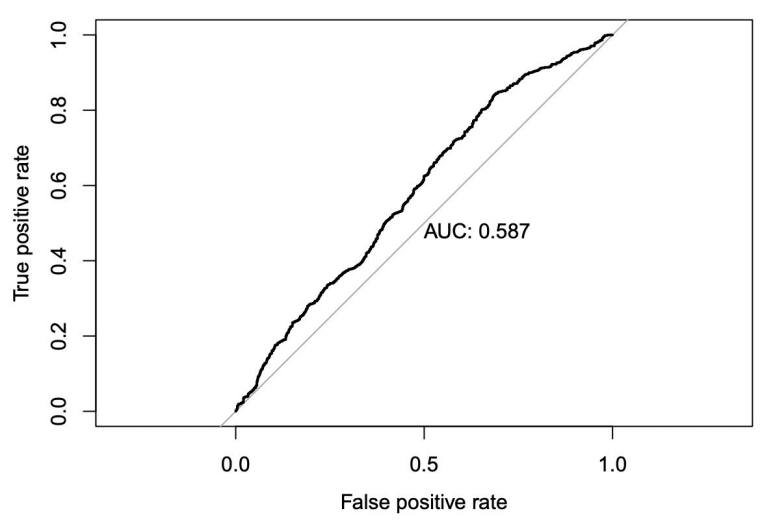
- Building a logistic regression model to classify the RRP as high or low (over/under median)
- starts from a full model and simplifies by a process of backwards elimination



Distribution of train-test sets and labels
(70%-30% train-test split)

Initial model (Using Seasons)

- Variables: demand, mean_temperature, solar_exposure, log(rainfall +1), holiday, school_day, season
- In the first step, it was decided to eliminate solar exposure (because of the high p-value) and holidays (because it was very unbalanced).



ROC of the first model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.500e+00	6.884e-01	-3.632	0.000281
demand	1.368e-05	4.894e-06	2.796	0.005178
mean_temperature	3.075e-02	1.696e-02	1.813	0.069895
solar_exposure	-4.516e-03	9.735e-03	-0.464	0.642695
log(rainfall + 1)	-1.027e-01	7.209e-02	-1.425	0.154113
holidayTRUE	-3.908e-01	3.321e-01	-1.177	0.239283
school_dayTRUE	3.668e-01	1.191e-01	3.079	0.002075
seasonspring	3.330e-01	1.806e-01	1.844	0.065232
seasonwinter	3.771e-01	2.529e-01	1.491	0.135993
seasonautumn	2.616e-01	1.918e-01	1.364	0.172637

- Null Deviance: 2011.5
- Residual Deviance: 1969.1

Second and third model

- Second Model: demand, mean_temperature, school_day, season
- Third Model: demand, school_day, season

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.790e+00	6.296e-01	-4.431	9.36e-06
demand	1.529e-05	4.727e-06	3.234	0.00122
mean_temperature	2.934e-02	1.669e-02	1.758	0.07873
log(rainfall + 1)	-9.257e-02	7.058e-02	-1.312	0.18967
school_dayTRUE	3.838e-01	1.181e-01	3.250	0.00115
seasonspring	3.589e-01	1.777e-01	2.020	0.04342
seasonwinter	4.150e-01	2.345e-01	1.770	0.07674
seasonautumn	2.978e-01	1.692e-01	1.761	0.07831

- Null Deviance: 2011.5
- Residual Deviance: 1970.7

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.936e+00	6.201e-01	-4.734	2.2e-06
demand	1.537e-05	4.726e-06	3.251	0.00115
mean_temperature	3.437e-02	1.625e-02	2.115	0.03441
school_dayTRUE	3.828e-01	1.180e-01	3.244	0.00118
seasonspring	3.767e-01	1.772e-01	2.127	0.03345
seasonwinter	4.468e-01	2.331e-01	1.917	0.05527
seasonautumn	3.077e-01	1.689e-01	1.822	0.06846

- Null Deviance: 2011.5
- Residual Deviance: 1972.4

Fourth and fifth model

- Fourth Model: demand, school_day, season
- Fifth Model: demand, school_day

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.361e+00	5.528e-01	-4.271	1.95e-05
demand	1.655e-05	4.679e-06	3.536	0.000406
school_dayTRUE	3.891e-01	1.178e-01	3.303	0.000957
seasonspring	1.996e-01	1.554e-01	1.285	0.198781
seasonwinter	9.718e-02	1.644e-01	0.591	0.554474
seasonautumn	1.622e-01	1.536e-01	1.056	0.290967

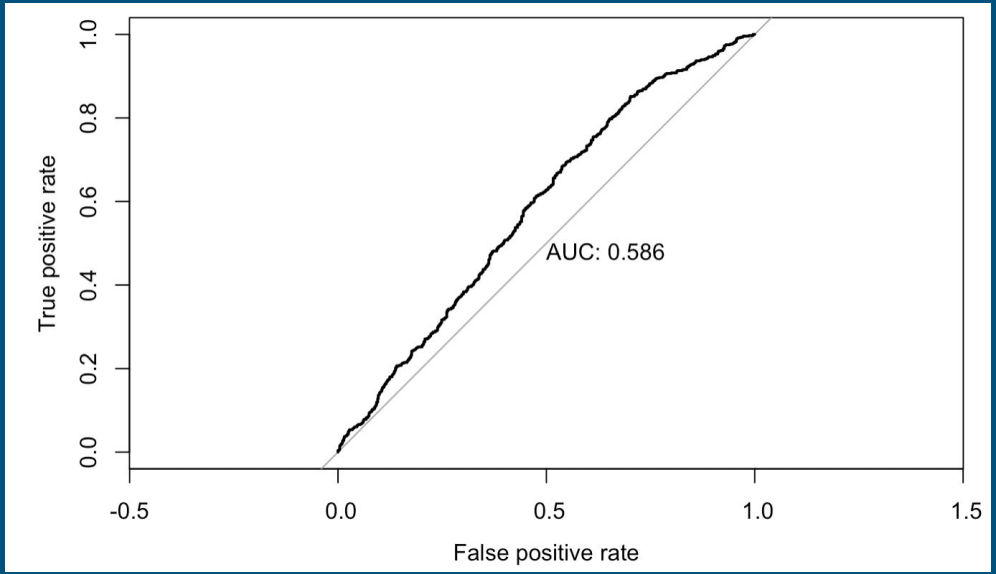
- Null Deviance: 2011.5
- Residual Deviance: 1976.9

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.165e+00	4.876e-01	-4.44	9e-06
demand	1.575e-05	4.092e-06	3.85	0.000118
school_dayTRUE	4.125e-01	1.165e-01	3.54	0.000400

- Null Deviance: 2011.5
- Residual Deviance: 1978.8

Model Selection (Using Seasons)

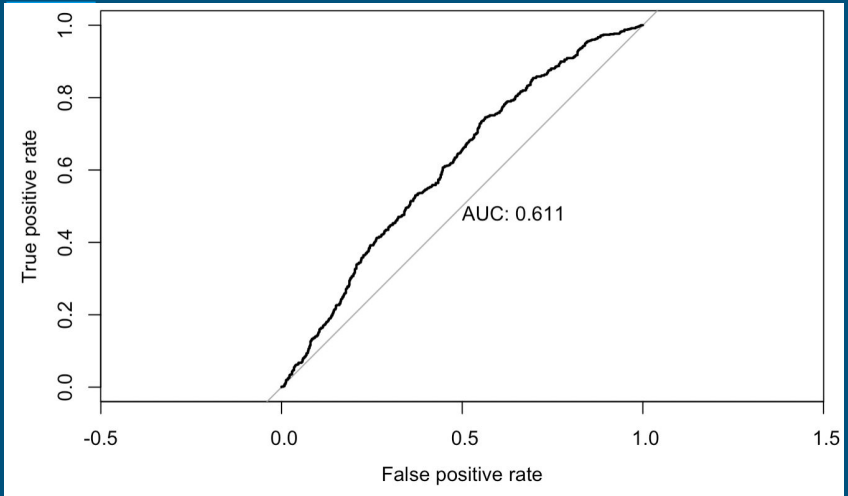
- In this part we are left with the third model, due to the change of the deviance
- Third Model: demand, school_day, season



ROC of the third model

Initial model (Using Months)

- Variables: demand, mean_temperature, solar_exposure, log(rainfall +1), holiday, school_day, month



ROC of the first model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.566e+00	7.374e-01	-3.480	0.000501
demand	1.179e-05	5.057e-06	2.332	0.019676
mean_temperature	4.355e-02	1.929e-02	2.258	0.023958
solar_exposure	1.861e-03	1.059e-02	0.176	0.860513
log(rainfall + 1)	-8.899e-02	7.371e-02	-1.207	0.227347
holidayTRUE	-4.682e-01	3.368e-01	-1.390	0.164468
school_dayTRUE	5.824e-01	1.461e-01	3.985	6.74e-05
month2	-4.967e-01	2.980e-01	-1.667	0.095495
month3	-2.519e-01	2.973e-01	-0.847	0.396806
month4	4.299e-01	3.072e-01	1.399	0.161746
month5	1.738e-01	3.532e-01	0.492	0.622572
month6	4.707e-01	3.891e-01	1.210	0.226334
month7	6.300e-01	3.816e-01	1.651	0.098750
month8	-5.958e-02	3.642e-01	-0.164	0.870078
month9	2.631e-01	3.176e-01	0.828	0.407504
month10	9.959e-02	2.979e-01	0.334	0.738114
month11	2.787e-02	3.115e-01	0.089	0.928710
month12	-2.330e-01	2.962e-01	-0.787	0.431539

- Null Deviance: 2011.5
- Residual Deviance: 1954.1

Second and third model

- Second Model: demand, mean_temperature, school_day, month
- Third Model: demand, school_day, mean_temperature

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.888e+00	6.523e-01	-4.427	9.56e-06
demand	1.310e-05	4.949e-06	2.647	0.00812
mean_temperature	4.987e-02	1.873e-02	2.663	0.00774
school_dayTRUE	5.959e-01	1.456e-01	4.093	4.25e-05
month2	-4.635e-01	2.954e-01	-1.569	0.11666
month3	-2.366e-01	2.871e-01	-0.824	0.40986
month4	4.080e-01	2.797e-01	1.459	0.14456
month5	2.040e-01	3.153e-01	0.647	0.51777
month6	4.969e-01	3.517e-01	1.413	0.15779
month7	6.740e-01	3.460e-01	1.948	0.05144
month8	-2.120e-02	3.377e-01	-0.063	0.94994
month9	3.220e-01	3.007e-01	1.071	0.28430
month10	1.577e-01	2.921e-01	0.540	0.58937
month11	5.552e-02	3.097e-01	0.179	0.85772
month12	-2.102e-01	2.944e-01	-0.714	0.47521

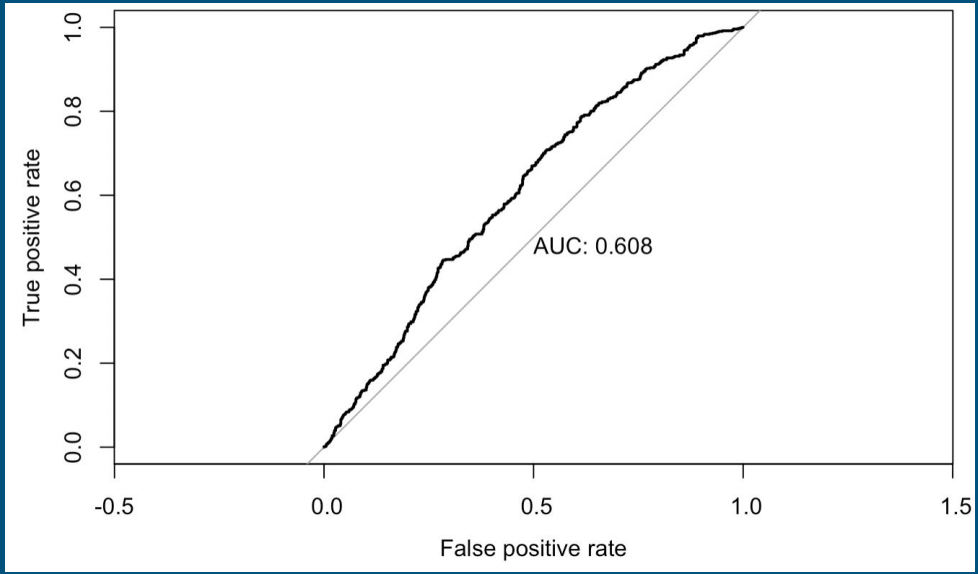
- Null Deviance: 2011.5
- Residual Deviance: 1957.7

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.456e+00	5.630e-01	-4.363	1.28e-05
demand	1.657e-05	4.172e-06	3.971	7.15e-05
mean_temperature	1.195e-02	1.142e-02	1.046	0.295445
school_dayTRUE	4.180e-01	1.167e-01	3.582	0.000342

- Null Deviance: 2011.5
- Residual Deviance: 1977.7

Model Selection (Using Months)

- In this part we are left with the second model, due to the change of the deviance
- Second Model: demand, mean_temperature, school_day, month



ROC of the second model

Model Evaluation (Linear Regression)

To evaluate the performance of our models, we get the MSE (Mean Squared Errors) of the test set by doing a simple test-train-split and compare it to the MSE obtained by 10-fold cross validation. We assign the 80 percent of the data to the training set and the rest of the data to the test set.

- MSE without cross validation: 80388633
- MSE with cross validation: 88238532

##	folds	cross.validation.errors
## 1	1	82540383
## 2	2	84387844
## 3	3	94685089
## 4	4	89769280
## 5	5	92233213
## 6	6	85484240
## 7	7	96069217
## 8	8	81960217
## 9	9	91930535
## 10	10	83325300

Model Evaluation (Logistic Regression)

We show the evaluation of the second model using months. The evaluation was carried out on the test set (623 samples)

logistic.pred	FALSE	TRUE
FALSE	143	103
TRUE	170	207

Confusion Matrix for predicted RRP with the model selected

- Accuracy: 56%
- Specificity: 45.69%
- Sensitivity: 66%

Using the threshold of 0.43 (the best for this model).

Conclusion

- For the regression model to predict the demand, the best results are obtained by considering the "solar_exposure", "mean_temperature", "seasons" and "school_days" variables. The behavior of the regression models had non linear correlation with two variables.

- While performing the hypothesis test on variances in the EDA section, we could not find evidence that the variances were equal, because the data was very unbalanced.

- For the logistic regression the model that obtained the best results used the variables: "demand", "mean_temperature", "school_day", "month", and reported an area under the ROC curve equal to 0.590. The best threshold for this method was 0.43 and it is possible to see that the accuracy is quite low (55%).

Thank you for your attention.

Any questions?