Università degli studi di Milano
Data Science for Economics

Statistical Learning

Bike Sharing in Seoul

by
Jan Philip Richter
Matriculation Number: 20547A

submitted to
Prof. Dr. Silvia Salini

September 11, 2023

Abstract

This report covers the analysis of bike sharing data in the South Korean capital Seoul. The goal is to investigate which relationships exist between the weather conditions and the number of rented bikes at a given time. Furthermore the goal is to investigate which statistical modelling techniques yield desirable results in predicting the number of rented bikes.

Unsupervised machine learning techniques are used to graphically analyse the potential effect of the predictors on the bike sharing demand and to explore potential patterns in the data set. Using Principal Component Analysis the dimensionality of the data set is reduced and with a clustering algorithm the data is split into groups with different characteristics that give a strong insight about the effects the weather conditions have on the number of rented bikes.

A variety of different statistical learning algorithms are then used to quantify the potential relation between the predictors and the demand of shared bikes. The advantages and disadvantages of applying those methods on the data set are discussed and the performance of the different machine learning algorithms is then compared and analysed.

Contents

1 Introduction

With over 9 million inhabitants, Seoul is one of the largest cities in the world and thus faces a huge demand in both public and individual transportation. With the goal of reducing road traffic the city has introduced a public bike rental service, called *Ttareungyi*, in 2015, whose demand has increasingly risen since, with the number of total bicycles going up from 20,000 in 2018 to over 37,000 in 2021 [3].

1.1 Data set

The data set used for the statistical analysis was originally obtained from the SEOUL OPEN DATA PLAZA [6] and is available for download at the Machine Learning Repository of the University of California [4].

The data set contains 8,760 observations, over a time span of exactly one year ranging from 01.12.2017 until 30.11.2018. Each observation represents exactly one hour of the day and holds information about the number of bikes that were rented at that specific hour as well as information about different weather conditions at the current time.

The variables contained in the data set are shown in the table below:

Name	**Description**	**Type**	**Measurement**
Date	The date of the observation	Datetime	dd/mm/yyyy
Rented Bike Count	Number of rented bikes in that hour	Integer	0,1,2 ... 3556
Hour	The hour of the day	Integer	0,1,2 ... 23
Temperature	Temperature in degrees Celsius	Continuous	$°C$
Humidity	Air humidity in percent	Integer	%
Wind Speed	Wind speed in Metres per second	Continuous	m/s
Visibility	Visibility in 10 Metres	Integer	$10m$
Dew Point Temperature	Dew point temperature in degrees Celsius	Integer	$°C$
Solar Radiation	Radiation in Mega Joule per square metre	Continuous	MJ/m^2
Rainfall	Rainfall in Millimetres	Integer	mm
Snowfall	Snowfall in Centimetres	Integer	cm
Seasons	Season of the year	Categorical	Spring, Summer, Autumn, Winter
Holiday	Indicator if the day was a public holiday	Binary	Holiday / No Holiday
Functioning Day	Indicator if the bike service was available	Binary	Yes / No

1.2 Problem description

The goal of this report is to explore the patterns and connections in the data set and between the variables and give possible explanations for the phenomena found.

Furthermore the goal is to inspect how the number of rented bikes at a given hour may be effected by the other variables in the data set. The aim is to use various

statistical modeling methods to predict the target variable *Rented Bike Count* using the variables about the weather condition and time information as predictors, quantifying and interpreting the individual effects on the number of bikes rented and providing possible explanations and intuitions to the findings.

Moreover it will be investigated how the different approaches vary in predictive accuracy on unseen data, how they perform under the common problem of multicollinearity, as well as the advantages and disadvantages between the modeling methods.

2 Data Preparation

Before beginning with the application of any unsupervised or supervised statistical learning methods the data has to be cleaned, prepared and if necessary transformed.

Multiple steps of preparation are taken to enhance the usability of the data for further analysis and modeling.

2.1 Data preprocessing

Data Classes

The variables data classes were changed so that every numeric feature is either *int* or *dbl*, the categorical features are converted to *factors* and the date is converted to *datetime* format.

Missing Values

The data set has been checked for missing values, infinite values and duplicates but none were found.

Removing Functioning Day

The variable *Functioning Day* is an indicator variable which shows if the bike service was available for usage at a given time or not.

For all observations were *Functioning Day* = NO, the number of rented bikes was obviously 0. As there is no information contained in the variable about the number of bikes rented, when the service is functioning and as it is pointless to analyse observations, were the rental service is not even available, all 295 observations where the service was not up for usage were removed.

The problem trivially changes to now predict the amount of bikes rented, given that the bike sharing service is available.

Renaming

The variables have been renamed for simplicity in the further analysis.

New variable names:

- Date → date
- Rented Bike Count → count

- Hour → hour
- Temperature → temperature
- Humidity → humidity
- Wind Speed → wind_speed
- Visibility → visibility
- Dew Point Temperature → dp_temperature
- Solar Radiation → solar _radiation
- Rainfall → rainfall
- Snowfall → snowfall
- Seasons → seasons
- Holiday → holiday

2.2 Descriptive Statistics

Now that the data set is ready for usage, the first step is to start with a descriptive analysis of the variables.

2.2.1 Correlation Analysis

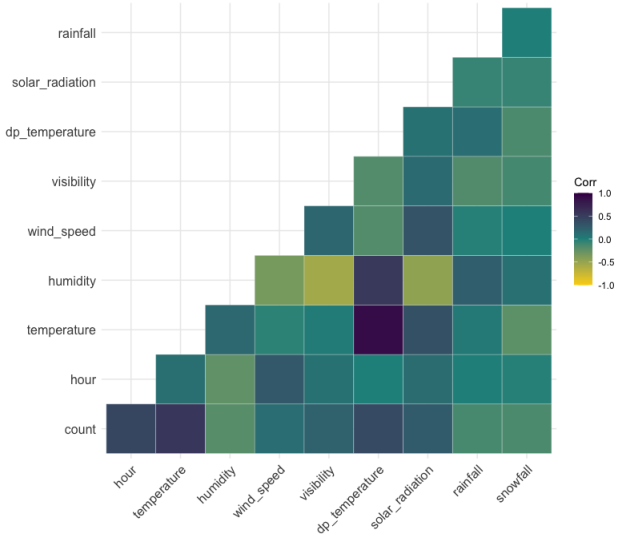
An analysis of the correlation between the numeric variables allows us to explore potential relationships between the variables. The temperature, hour of the day and the dew point temperature seem to have the strongest relationship with the number of bikes rented.

With a correlation coefficient of 0.91 there also appears to be a strong connection between the temperature and the dew point temperature. The latter also shows a relatively strong correlation with humidity. In fact, the dew point temperature can be almost perfectly approximated using the so called *Magnus formula* [5], so there exists almost perfect multicollinearity between the dew point temperature and the regular temperature and humidity.

Based on this knowledge it would be reasonable to exclude dew point temperature at this point already from any further analysis but we intentionally keep it in the data set to observe the effects of this relationship on the models which will be introduced later on in the analysis.

The correlation matrix also gives an indication if the variables have a positive or a negative effect on the number of bikes rented. While *temperature* is positively correlated with *count*, *humidity* is negatively correlated.

Figure 2.1: Correlation between the variables



2.2.2 Exploratory Data Analysis

As an exemplary variable we will only focus on the analysis of the *wind_speed* variable. A more in-depth analysis on all of the variables can be found in the respective Github repository [7].

Summary statistics

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0	0.9	1.5	1.75	2.3	7.4

The summary statistics show that there are no implausible values for the wind speed in the data set. The maximum value of $7.4 m/s$ is classified as a *moderate breeze* on the Beaufort-scale [2]. (For all of the variables in the data set the range of values was checked for plausibility, but no implausible values were found.)

We can also assume that the distribution has some skewness, as the mean is higher than the median and the maximum value is roughly three times higher than the 3rd quartile.

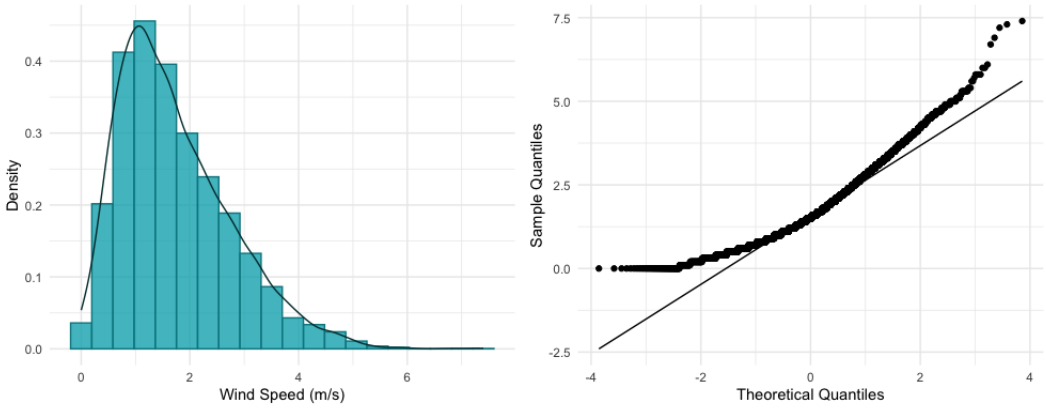
Distribution

The histogram confirms the assumption that the distribution is right-skewed. The estimated moment coefficient of skewness is 0.89 and also the quantile-quantile plot indicates, that the Wind Speed variable is not normally distributed.

For most statistical modeling methods, it is not a requirement, that the predictors or the target variable have to be normally distributed, however it is still undesirable

using skewed variables, as this often results in non-normal residuals, which compromises the inference drawn from the model. The problem of skewness is thus tackled in the following section.

Figure 2.2: Histogram with Kernel Density Estimation and QQ-Plot for Wind Speed



2.3 Feature engineering and data transformation

As some of the variables are not properly suited for usage in statistical modeling, due to their distributions or range, so it is useful to perform feature engineering and transformations to enhance the usability, distributional properties or predictive qualities.

The transformations and changes made are briefly explained in the following.

Count

The *count* variable, which represents the number of bikes rented in an hour, will be log-transformed, as it is right-skewed.

After the transformation the variable now has a small-left skew, but an overall distribution which is closer to normal than before, has a smaller absolute skewness and lower variance, which will increase the predictive quality of the statistical models.

The new variable is denoted *log_count*.

Hour

The *hour* variable indicates the hour of day and is potentially a useful predictor for the number of rented bikes. Nonetheless it has a fundamental flaw, which makes it unusable in the current form, which is the range that goes from 0 to 23. The problem is that observations with $hour = 23$ are very close to observations with $hour = 0$ in terms of time distance, but as far apart as possible in terms of absolute value.

Using one-hot or target-encoding is not a suitable method of usage in this case as the measure of distance between the values gets completely lost this way.

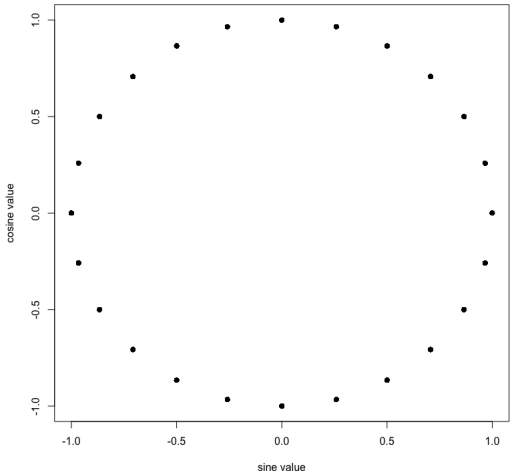
An appropriate way to encode this variable for further usage is cyclical encoding [1]. In this case sine-cosine-encoding is used to generate two new predictors from the hour of day:

$$x_{sin} = sin(\frac{2\pi * hour}{24}) \tag{2.1}$$

$$x_{cos} = cos(\frac{2\pi * hour}{24}) \tag{2.2}$$

The graphical representation of our new features for the time of day now resembles a clock with the 24 different possible times, with adjacent times now being close to one another in geometrical distance.

Figure 2.3: Cyclical encoding of the hour of the day



The two new variables are denoted *sin_hour* and *cos_hour*

Date

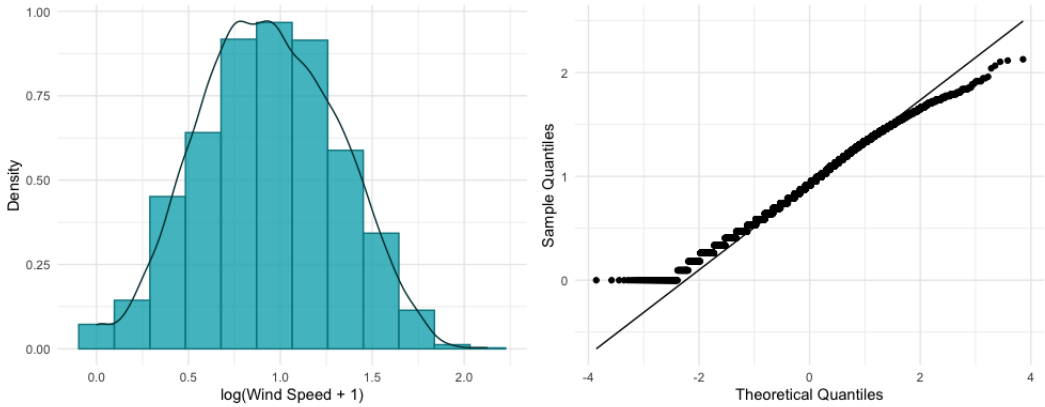
From the *date* variable we extract the day of the week of that day and use the same sine-cosine-encoding method as previously to obtain two new predictors, denoted *sin_dow* and *cos_dow*.

Wind Speed

As discussed in the previous section, the *wind_speed* variable will be transformed to obtain a closer-to-normal distribution.

A log transformation is appropriate because of the right-skewness, but as there are observations where *wind_speed* = 0, a *log1p-transformation* will be performed, where every value of *wind_speed* gets added by a constant of 1 before logarithmising.

Figure 2.4: Histogram with Kernel Density Estimation and QQ-Plot for log(Wind Speed + 1)



The distribution after the transformation is now way less skewed (estimated moment coefficient of skewness is now 0.01)and more appropriate for usage as a predictor in a statistical model.

The new variable is denoted *log_wind_speed*.

Visibility

The *visibility* variable is extremely left-skewed and is thus square-root-transformed. Unfortunately the distribution stays left-skewed even after the transformation, however with an overall lower variance. This may have some negative implications on the statistical modeling later on.

The new variable is denoted *sqrt_visibility*.

Rainfall and Snowfall

The variables *rainfall* and *snowfall* are both extremely right-skewed and also stay that way after a log-transformation. This is due to the fact, that the absolute majority of observations did not record any rain- or snowfall. As mentioned before this may have negative consequences in the modeling process later on.

The new variables are denoted *log_rainfall* and *log_snowfall*.

Alternatively the variables could have been one-hot encoded into *no rainfall / rainfall* with the downside of loosing all information about a high or low amount of rain- or snowfall.

Holiday and Seasons

The variables *holiday* and *seasons* get one-hot encoded to obtain a dummy-variable for every category, to be used in the statistical models.

3 Unsupervised Learning

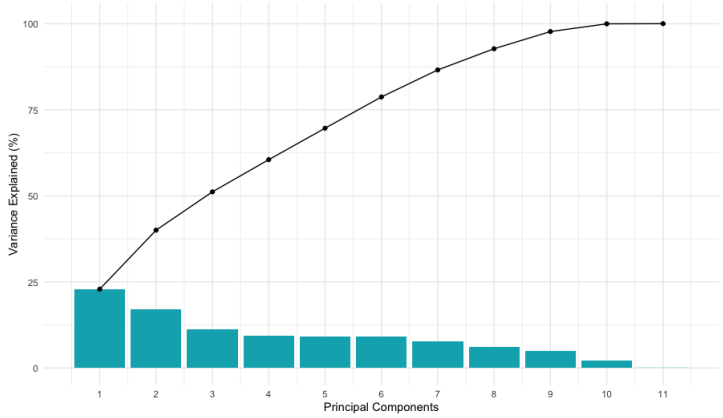
This chapter takes use of two unsupervised machine learning techniques, namely principal component analysis and k-means clustering, with the aim of exploring different characteristics in the data and relationships between the variables

3.1 Principal Component Analysis

The Principal Component Analysis allows us to investigate how much if the variability in the data set can be explained by a reduced dimensionality and to graphically analyse potential patterns and relationships in the data.

For the calculation all numeric variables, except for count, were used. The data was scaled, centred and the principal components have been calculated using singular vector decomposition. This method has the advantage of higher numerical accuracy compared to Eigendecomposition of the covariance matrix.

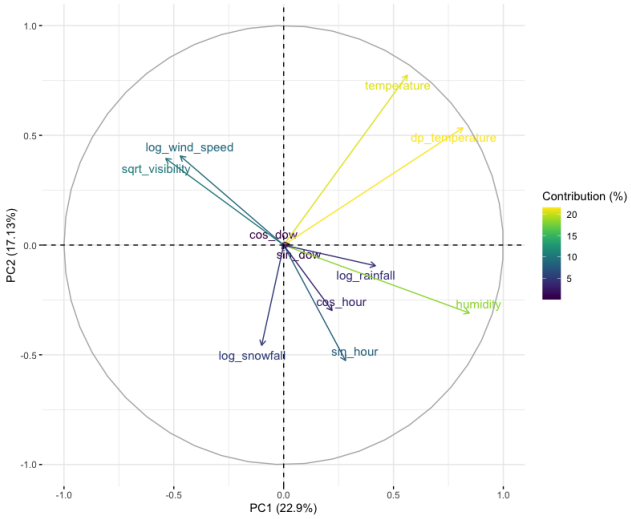
Figure 3.1: Variance explained by the principal components



As can be seen in Figuer 3.1, the dimensionality cannot be significantly reduced without losing much information, as even the 9th out of 11 principal components still explains roughly 5% of the overall variance in the data set.

We will continue to focus on the first 2 principal components, that explain a cumulative 40% of the overall variance, which is not enough to capture the majority of information in the data but might still allow us to capture some patterns or relationships. Additionally only using the first 2 principal components, makes a graphical analysis more comprehensible and straight forward.

Figure 3.2: Correlation circle of the first 2 principal components



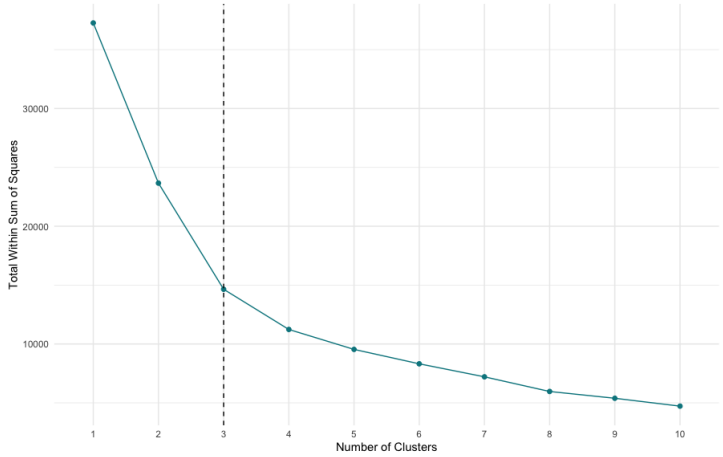
Temperature and dew point temperature yield the highest contribution to the first two principal components but are not significantly correlated with either one of them. Also for the other variables there is not an obvious pattern of correlation with the first 2 principal components to be observed. This means that there is no intuitive explanation of the information displayed by those 2 principal components but they are still be useful to further analyse patterns in the data with clustering methods.

3.2 Clustering

The first 2 principal components will be used partition the data into clusters and analyse the potential patterns that may be observed.

The k-means algorithm will be used to perform the clustering. There is not an intuitive number of clusters that are desirable to obtain, as our target variable for the regression modeling is continuous and there are no fixed number of categories to classify. We thus graphically analyse the reduction of total within sum of squares each higher number of clusters yields to determine an appropriate number of clusters.

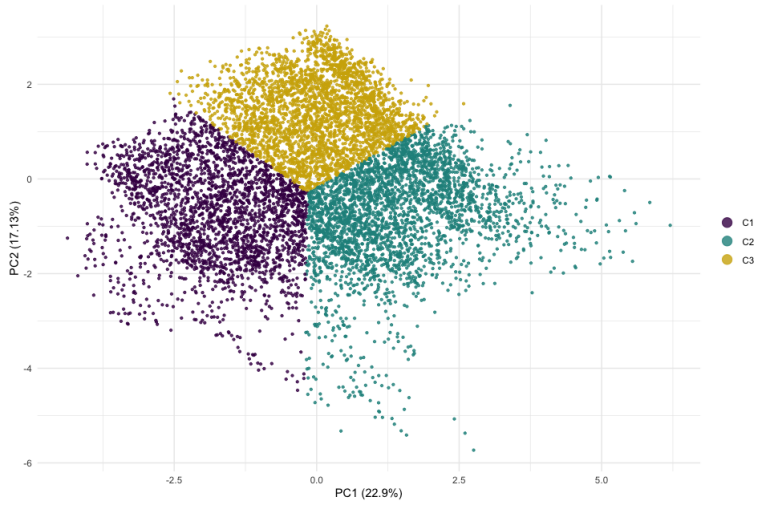
Figure 3.3: Optimal number of clusters for k-means



Moving from 2 to 3 clusters significantly reduces the within sum of squares while 4 clusters or above does not yield as high of a reduction in within sum of squares anymore. Therefore, partitioning the data into 3 clusters appears to be reasonable.

To generate a stable partitioning of the data we run the algorithm with 1,000 initialisations of the 3 cluster centers.

Figure 3.4: 3-Means Clustering with first 2 Principal Components

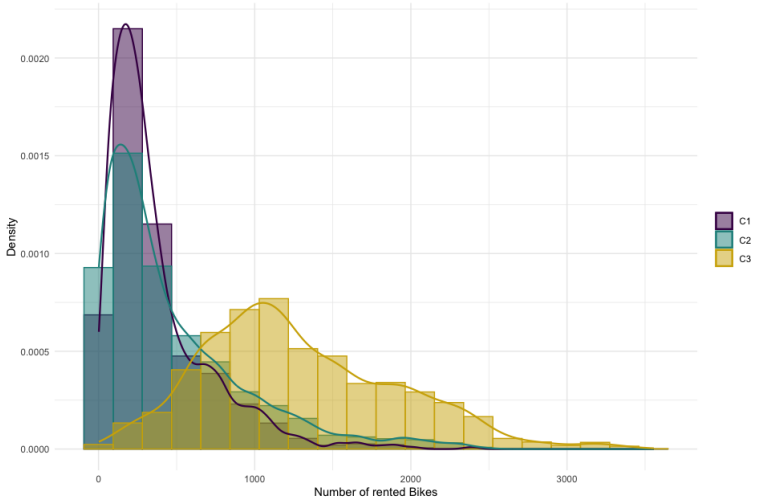


The scatterplot shows that there is one large point cloud with some extreme points especially on the right and bottom left. Those extreme value observations could very well have some distinct properties while in the middle of the point cloud the partitioning does not seem to imply a vast difference between the 3 clusters.

The absence of a clear partitioning in the data is not an undesirable result however, as a homogeneous population could potentially improve the quality of the regression analysis.

To further analyse the differences and patterns in the clusters we can observe the characteristics of the data points in each cluster.

Figure 3.5: Histograms and Kernel Density Estimations for each cluster

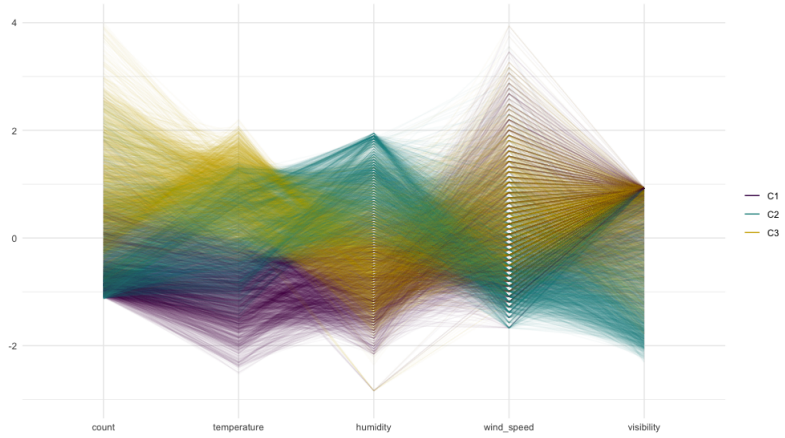


The distribution of the number of rented bikes for each cluster shows that there are some observable differences between the clusters.

Especially cluster 3 seems to consist of data points that appear to represent a significantly higher number rented bikes. Clusters 1 and 2 however follow a fairly similar distribution, regarding the number of bikes rented, with the main difference being, that cluster 1 has a higher peak in the density for relatively low numbers of bikes rented.
(Note: The number of bikes on the x-axis shows the non-transformed variable for better interpretability)

The parallel coordinates plot provides some further insights into the patterns each cluster conveys:

Figure 3.6: Parallel Coordinates for each Cluster



Cluster 1

The overall pattern indicates predominantly low temperature and low humidity data points. The wind speed is also higher on average than for the other clusters. This cluster may therefore potentially represent cold and windy hours.

Cluster 2

Even though the count for cluster 1 and 2 seems to follow a rather similar distribution, we can see that the weather condition of the data points, which are represented by cluster 2 are vastly different. Cluster 2 is mainly made up of fairly high temperature observations, paired with extremely high humidity and extremely low wind speeds and low visibility. This cluster may therefore represent hours where the air was extremely thick and smog levels were potentially high.

Cluster 3

The data points contained in this cluster generally have a high temperature, paired with average humidity and wind speed levels and average to high visibility levels. This cluster may therefore represent the nice and pleasant hours.

(Note: Only a subset of the covariates used for the principal component analysis are shown in the parallel coordinates plot and the variables count, wind speed and visibility represent the non-transformed versions for better intepretability of the cluster-characteristics.)

3.3 Insights

The cluster analysis provided some useful perceptions about the patterns and relationships in the data. The key-findings can be summarised as:

- Low temperatures paired with high wind speeds seem to be associated with a low number of rented bikes
- High humidity with low wind speeds also seem to be associated with a low number of rented bikes
- High temperatures paired with average humidity and wind speeds associated with a high number of rented bikes

The clustering helped not only to indicate the different patterns in the data points but also to already provide some useful insights on the relationship between the weather conditions and the number of bikes that are rented. This is quite a remarkable result, given the fact, that only the first 2 principal components were used to partition the data. The first two principal components explained only 40% of the overall variance in the data and contained no information about the categorical variables or the target variable.

In the next chapter we will quantify the relationship between the weather conditions and categorical variables and the number of bikes rented through the use of multiple regression techniques.

4 Supervised Learning

To quantify the individual effects of the explanatory variables on the target variable *log_count* we will use a variety of different regression models.

Those models will be fit on a training set, which consists of 70% of the data (5,910 observations) and tested for predictive accuracy on the unseen data of a test set, made up from the remaining 30% of the data (2,555 observations).

4.1 Parametric Modeling

Parametric regression methods have the benefit of simplicity and interpretability. The regular linear regression with ordinary least squares parameter estimation will be compared to regulisation methods in the light of the multicollinearity issues previously discussed.

4.1.1 Linear Regression

Principal Component Regression

The first model we consider makes use of the first two principal components that were computed previously and takes the following form:

$$\widehat{log\_count}_i = \hat{\beta}_0 + \hat{\beta}_1 PC1_i + \hat{\beta}_2 PC2_i \tag{4.1}$$

The model yields an R^2 value of 0.45 which is an impressive result, considering the model is extremely sparse with only two regressors. Espescially given the fact, that the first two principal components were computed from the numeric variables only and explained just 40% of the variance in the data.

Figure 4.1: Principal Component Regression Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.085517   0.011078 549.314  < 2e-16 ***
pc1         -0.025483   0.006947  -3.668 0.000247 ***
pc2          0.566877   0.008062  70.318  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8517 on 5907 degrees of freedom
Multiple R-squared:  0.4562,    Adjusted R-squared:  0.456
F-statistic:  2477 on 2 and 5907 DF,  p-value: < 2.2e-16
```

There is not much interpretation to provide however, as there was no intuitive explanation of the information found for the principal components.

Multiple Linear Regression

The next model we fit takes use of all possible variables as regressors to explain *log_count*:

$$\begin{aligned}\widehat{log\_count}_i =& \hat{\beta}_0 + \hat{\beta}_1 temperature_i + \hat{\beta}_2 dp\_temperature_i + \hat{\beta}_3 humidity_i + \\ & \hat{\beta}_4 log\_wind\_speed_i + \hat{\beta}_5 solar\_radiation_i + \hat{\beta}_6 sin\_hour_i + \\ & \hat{\beta}_7 cos\_hour_i + \hat{\beta}_8 sin\_dow_i + \hat{\beta}_9 cos\_dow_i + \hat{\beta}_{10} log\_rainfall_i + \\ & \hat{\beta}_{11} log\_snowfall_i + \hat{\beta}_{12} sqrt\_visibility_i + \\ & \hat{\beta}_{13} spring_i + \hat{\beta}_{14} summer_i + \hat{\beta}_{15} winter_i + \hat{\beta}_{16} holiday_i\end{aligned} \tag{4.2}$$

The baseline season for this model is autumn, which is why it is not included as one of the predictors. With an R^2 of 0.67 we have significantly improved the regressands variance explained by the model, compared to the principal component regression. This is of course to be expected, as we include all 16 possible predictors now. The resulting model indicates, that every regressor is significant on a significance level of $\alpha = 5\%$, except for *sqrt_visibility*.

There are a couple of interesting phenomena in this model. Despite the almost perfect multicollinearity between *dp_temperature* and *temperature* and *humidity*, all three covariates have a statistically significant p-value. Without having knowledge about the underling multicollinearity one might therefore assume, that those variables do not contain the same information. However we can observe, that despite the statistical significance of all of these three variables we obtained a very unpleasant side effect. The coefficient of *temperature* is negative, which implies that a higher temperature leads to a lower amount of bikes rented. In this special log-level relationship the appropriate interpreta-

Figure 4.2: Regression Output with all Variables

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.358186   0.188051  44.446  < 2e-16 ***
temperature     -0.018444   0.007053  -2.615  0.00895 **
dp_temperature   0.060272   0.007347   8.203 2.84e-16 ***
humidity        -0.027192   0.001984 -13.704  < 2e-16 ***
log_wind_speed  -0.125777   0.027442  -4.583 4.67e-06 ***
solar_radiation -0.130756   0.018506  -7.065 1.79e-12 ***
sin_hour        -0.511930   0.014969 -34.200  < 2e-16 ***
cos_hour        -0.245556   0.018372 -13.366  < 2e-16 ***
sin_dow         -0.096920   0.012165  -7.967 1.93e-15 ***
cos_dow         -0.058827   0.012358  -4.760 1.98e-06 ***
log_rainfall    -1.445776   0.033591 -43.041  < 2e-16 ***
log_snowfall    -0.094610   0.043874  -2.156  0.03109 *
sqrt_visibility -0.001321   0.001212  -1.090  0.27586
spring          -0.287388   0.026008 -11.050  < 2e-16 ***
summer          -0.206366   0.032225  -6.404 1.63e-10 ***
winter          -0.885846   0.036718 -24.126  < 2e-16 ***
holiday         -0.272679   0.041410  -6.585 4.95e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6617 on 5893 degrees of freedom
Multiple R-squared:  0.6725,    Adjusted R-squared:  0.6716
F-statistic: 756.3 on 16 and 5893 DF,  p-value: < 2.2e-16
```

tion would be, that a $1°C$ increase in temperature would reduce the estimated number of rented bikes by 1.8%. This is a highly implausible result, especially considering the fact, that *temperature* had a correlation coefficient of +0.56 with *count*. Also intuitively it would contradict the very reasonable assumption, that a higher temperature would likely increase the number of people willing to rent a bike.

For some of the variables there is no intuitive explanation to be provided for their effect on the estimated number of bikes rented. This is particularly the case for the cyclical variables. Nonetheless both predictors for the hour of day are extremely significant in estimating the target variable.

The most statistically significant variable in this model is the *log_rainfall* with a t-value of -43.04. Given the log-log relationship between regressor and regressand, the estimated coefficient states that a 1% increase in rainfall results in a 1.45% decrease in the predicted number of rented bikes, which seems like a very plausible effect.

Taking a look at the variance inflation factor also indicates, that there exists a severe problem of multicollinearity, especially with *temperature* and *dp_temperature*. Not only is the underlying multicollinearity likely the cause, for an increase in variance of the estimated coefficients, but it is also a probably explanation as to why we obtained an

extremely implausible estimator for the *temperature* variable.

Table 4.1: Variance Inflation Factors

temperature	**dp_temperature**	**humidity**
98.99	128.62	22.13

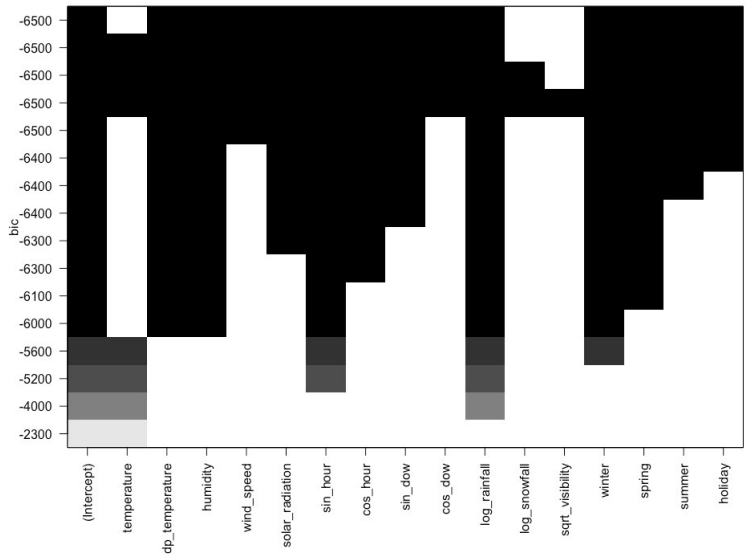
Best Subset Selection

The most straight forward way to tackle the problem of multicollinearity would be to manually exclude one of the problematic variables.

Alternatively a variable selection procedure might also yield a less problematic set of regressors, regarding the multicollinearity issue. The method applied in this case is the best subset selection procedure.

The unrestricted model contains 16 possible predictors, so there exist $2^{16} = 65,536$ possible combinations of those predictors. All of those possible models are fit and compared using the Bayesian Information Criterion (BIC) as an evaluation metric. The BIC is an appropriate metric in this case, as it penalises a higher number of regressors more severely than the AIC for instance.

Figure 4.3: Subset Selection with BIC



According to the BIC, the optimal subset of predictors excludes the variables *temperature*, *log_snowfall* and *sqrt_visibility*. This is a very pleasant result, as one of the troublesome predictors, namely *temperature* is now excluded from the model.

The regression output shows, that now all variables are highly significant, with the most significant variable still being *log_rainfall*, as before.

Figure 4.4: Subset Selection Regression Output

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.8852164  0.0544117 144.918  < 2e-16 ***
dp_temperature   0.0421659  0.0017246  24.450  < 2e-16 ***
humidity        -0.0226656  0.0007852 -28.865  < 2e-16 ***
log_wind_speed  -0.1280030  0.0273334  -4.683 2.89e-06 ***
solar_radiation -0.1446478  0.0177000  -8.172 3.67e-16 ***
sin_hour        -0.5047500  0.0146736 -34.399  < 2e-16 ***
cos_hour        -0.2474625  0.0182987 -13.524  < 2e-16 ***
sin_dow         -0.0960198  0.0121651  -7.893 3.49e-15 ***
cos_dow         -0.0582586  0.0122905  -4.740 2.19e-06 ***
log_rainfall    -1.4530239  0.0331099 -43.885  < 2e-16 ***
spring          -0.2783348  0.0253115 -10.996  < 2e-16 ***
summer          -0.2148825  0.0321329  -6.687 2.48e-11 ***
winter          -0.8714392  0.0356587 -24.438  < 2e-16 ***
holiday         -0.2722996  0.0414180  -6.574 5.30e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6622 on 5896 degrees of freedom
Multiple R-squared:  0.6718,    Adjusted R-squared:  0.6711
F-statistic: 928.4 on 13 and 5896 DF,  p-value: < 2.2e-16
```

Table 4.2: Variance Inflation Factors

dp_temperature	**humidity**
7.08	3.46

Also the variance inflation factors are now significantly lower, than previously and lie in an acceptable range of values.

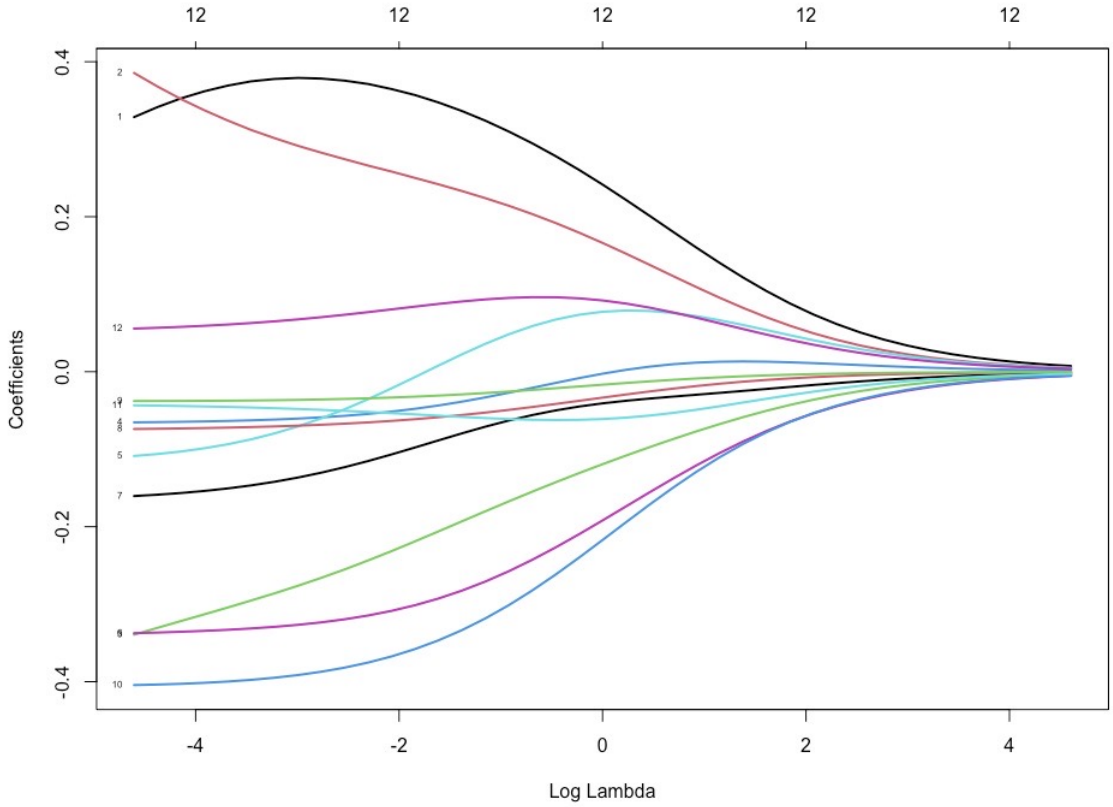
4.1.2 Shrinkage Methods

Another method of dealing with multicollinearity, opposed to the previously applied subset selection and variance inflation factors, are shrinkage methods.

Ridge Regression

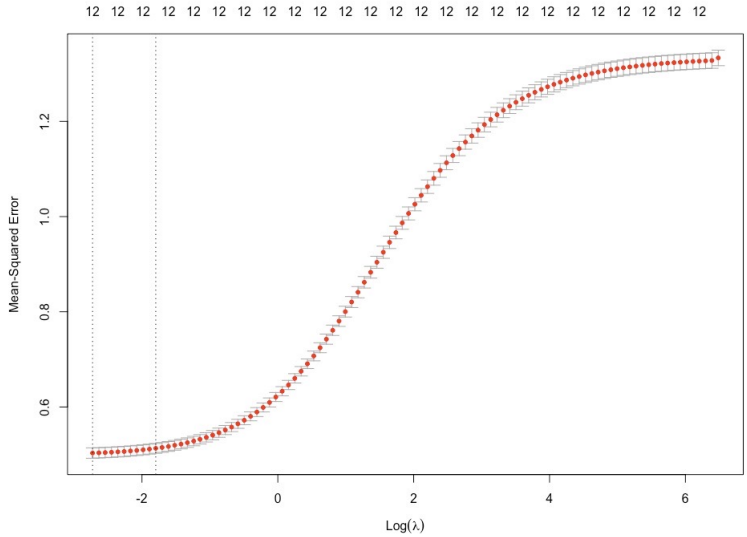
Ridge regression penalises the estimated coefficients based on their ℓ_2-norm. The ridge resgression penalises the coefficients for *dp_temperature* (nb. 2 - red line) and *humidity* (nb. 5 - light green line) the fastest, while the coefficient for temperature increases in the beginning. The issue of multicollinearity gets tackled, by significantly reducing the values of the coefficients for the troublesome covariates.

Figure 4.5: Ridge Regression Coefficients



To choose an adequate lambda, cross-validation was performed for a range of possible values. The metric for evaluation used is the mean squared error. The lambda which yields the lowest mean squared error is extremely low however, which is due to the fact that we have a large number of data points in our training set and thus, overfitting is not a huge issue. To obtain a higher penalisation, the lambda used for fitting the model is the highest lambda value within on standard error of the MSE minimising lambda. (In this case $\lambda = 0.2$)

Figure 4.6: Ridge Regression CV Lambda



In the end all variables stay in the model with the estimated coefficients shown below:

Figure 4.7: Ridge Regression Model Coefficients

(Intercept)	6.0879280610
temperature	0.3577651815
dp_temperature	0.2526660632
humidity	-0.2137699729
log_wind_speed	-0.0489724578
solar_radiation	-0.0006666005
sin_hour	-0.2949814623
cos_hour	-0.0953298868
sin_dow	-0.0606263698
cos_dow	-0.0306770447
log_rainfall	-0.3642634267
log_snowfall	-0.0601823485
sqrt_visibility	0.0646844950

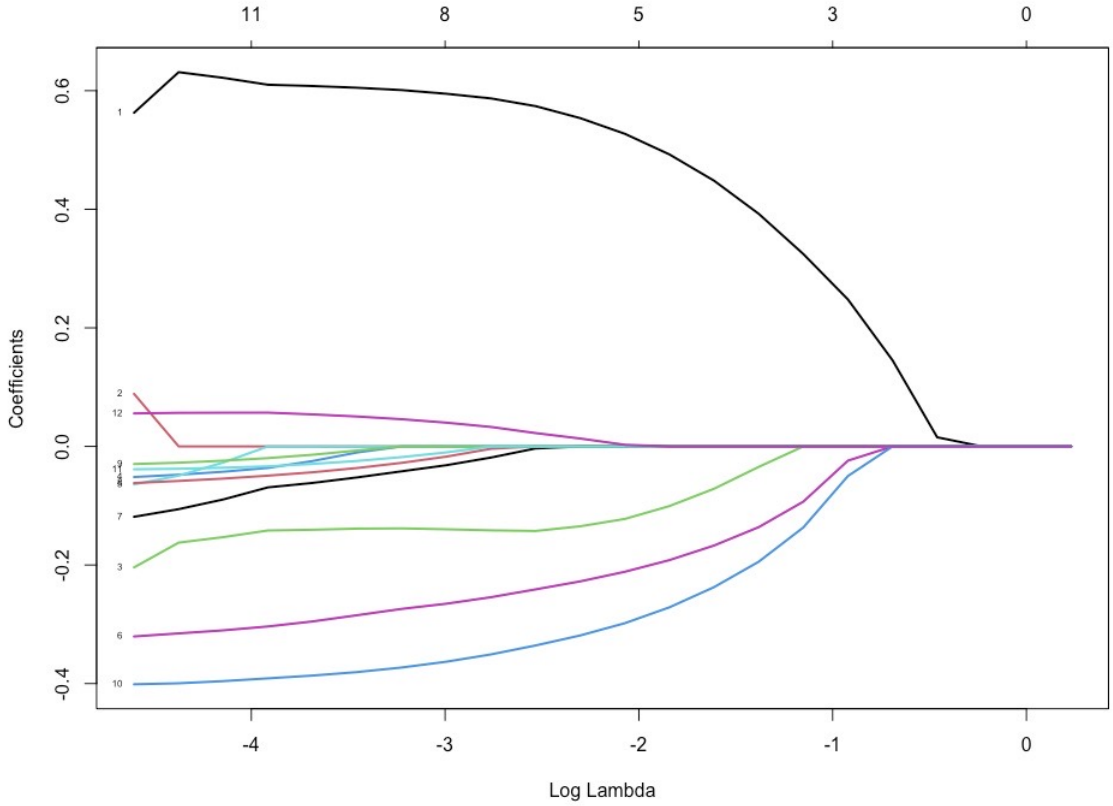
There is little interpretability to be given for the coefficients, as the data has been standardised before the estimation.

Lasso Regression

For the lasso regression, the ℓ_1-norm gets penalised, which can lead to variables being excluded from the model because their coefficients are estimated to be equal to 0.

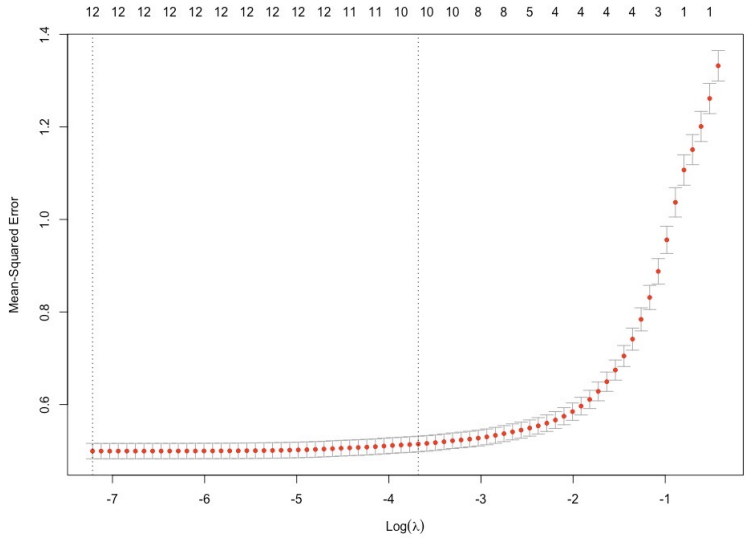
With an increasing shrinkage penalty, more and more variables get excluded from the model. The first variable to be excluded however is *dp_temperature*. It is also observable that *temperature* is deemed to be the most significant variable, as it is the last one to be excluded from the model and only with a fairly large shrinkage penalty.

Figure 4.8: Lasso Regression Coefficients



The same method of cross validation is performed, as before when using ridge regression. Again the lambda chosen for fitting the model will be the highest one, within one standard deviation of the MSE minimising lambda. The MSE minimising lambda is extremely low again, because of the large data set and will not lead to an exclusion of any variable. As the lasso regression is used to tackle the problem of multicollinearity however, a higher penalisation is seems appropriate. (In this case $\lambda = 0.028$)

Figure 4.9: Lasso Regression CV Lambda



The estimated coefficients from the lasso regression yield a more interesting result compared to ridge. Two variables are now excluded from the model, namely *dp_temperature* and *solar_radiation*. This is a desirable result, as *dp_temperature* was one of the troublesome variables regarding multicollinearity.

Figure 4.10: Lasso Regression Model Coefficients

```
(Intercept)      6.08792806
temperature      0.61786706
dp_temperature   .
humidity        -0.14080426
log_wind_speed  -0.02706887
solar_radiation  .
sin_hour        -0.29019762
cos_hour        -0.06193564
sin_dow         -0.04377084
cos_dow         -0.01274483
log_rainfall    -0.39654354
log_snowfall    -0.03387022
sqrt_visibility  0.02938638
```

Again the coefficients themselves do not have any straight forward interpretation.

4.2 Non-parametric Modeling

To account for potential non linear effects the usage of a generalised additive model (GAM) is suitable. The GAM was fit, smoothing all numeric variables except the cyclical variables.

Figure 4.11: GAM Coefficients

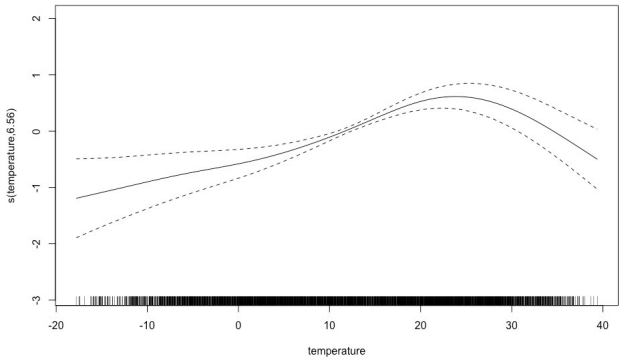
```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.35349    0.01910 332.603  < 2e-16 ***
sin_hour    -0.51631    0.01443 -35.774  < 2e-16 ***
cos_hour    -0.15656    0.02414  -6.485 9.61e-11 ***
sin_dow     -0.10167    0.01113  -9.135  < 2e-16 ***
cos_dow     -0.06999    0.01120  -6.249 4.43e-10 ***
spring      -0.26203    0.02391 -10.960  < 2e-16 ***
summer       0.01418    0.03418   0.415    0.678
winter      -0.80357    0.03970 -20.243  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                      edf Ref.df       F  p-value
s(temperature)     6.266  7.448  61.773  < 2e-16 ***
s(dp_temperature)  7.583  8.471   6.618  < 2e-16 ***
s(humidity)        7.391  8.201  62.072  < 2e-16 ***
s(log_wind_speed)  5.099  6.155   5.091 2.77e-05 ***
s(log_rainfall)    8.723  8.967 113.853  < 2e-16 ***
s(solar_radiation) 8.169  8.791  15.139  < 2e-16 ***
s(sqrt_visibility) 6.268  7.423   3.684 0.000445 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.736   Deviance explained = 73.9%
GCV = 0.35518  Scale est. = 0.35173    n = 5910
```

Interestingly all variables, except for the summer-dummy are now extremely significant. This indicates that there definitely seem to be some non-linear effects in the variables. As an example the smoothed *temperature* gives us some insight on the non-linear effect observable:

Figure 4.12: Smoothed effect of temperature



There seems to be a quadratic effect for the temperature, meaning that the predicted rented number of bikes first increase with increasing temperature and then decrease again, when the temperature gets too high.

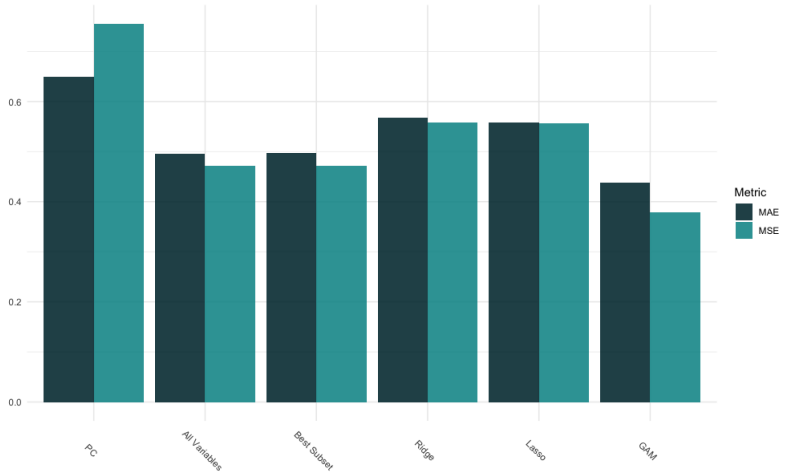
The goodness of fit of the GAM also increases naturally, as non-linear effects are now considered. With an $R^2 = 0.74$ and BIC = 11,046 this is the best model fit so far.

4.3 Model Evaluation

To evaluate the predictive accuracy of the different models introduced, their performance is tested on new and unseen test data. In this case the models will be used to predict the *log_count* of the 2,555 observations in the test set.

To evaluate the predictions the mean squared error (MSE) and the mean absolute error (MAE) are used.

Figure 4.13: Model Evaluations



The overall worst model in predictive accuracy is the one obtained from the principal component regression, which is to be expected, as it also had the lowest R^2 with a value of 0.45, out of all of the models and given the fact, that it only used the first two principal components as predictors.

The linear model using all predictors yields almost the exact same predictive quality as the model obtained from best subset selection. The subset selection model is therefore the preferred one, as it is sparser and has no obvious problems regarding multicollinearity anymore.

The two models using shrinkage methods also obtain almost the same MSE and MAE scores respectively, so between the two approaches the model coming from lasso regression might be the preferred one, as it excludes 2 variables completely. The shrinkage methods are not as good in terms of predictive quality, as the two linear regression models however, which might be down to the fact, that they did not contain any categorical variables as predictors, but only the numeric ones.

The best model in predictive accuracy is the GAM, which is down to the fact, that it considers also non-linear effects in the explanatory variables, of which there seemed to be some. It can therefore best explain the relationship between the regressors and the target variable.

5 Conclusion

The linear regression models yielded decent results in regards to interpretability as well as predictive accuracy. However it is easy to miss underlying multicollinearity, especially if there is a large training set to fit the model. This can lead to completely wrong estimations, as in the case of the *temperature* variable having a negative coefficient, even though there was statistical significance. In this case the best subset selection was able to remove the problematic predictors but there is no guarantee for that always being the case.

The shrinkage methods appear to be more reliable methods to tackle the problems of multicollinearity, especially when opting for a high shrinkage penalty. The down side is that almost all interpretability gets lost, compared to the standard linear regression, because of the standardisation of the variables. Only the direction of the effects, positive or negative, of the variables can be observed.

Performance-wise the GAM yielded the best results, as it is the only modeling technique among the ones discussed, that takes possible non-linear effects into consideration. The down side is again, that there is little interpretability of the smoothed variables, except for the possible graphical analysis.

Bibliography

[1] Anthony Adams and Peter Vamplew. Encoding and decoding cyclic data. *The South Pacific Journal of Natural Science*, 16(01), 1998.

[2] EL Delmar-Morgan. The beaufort scale. *The Journal of Navigation*, 12(1):100–102, 1959.

[3] Economic, Social Comission for Asia, and the Pacific. THE "TTAREUNGYI" PUBLIC BIKE SHARING SYSTEM IN SEOUL, 2021. URL https://www.unescap.org/sites/default/d8files/event-documents/Session%203_2.Seoul%20Public%20Bike%20Sharing.docx_.pdf.

[4] University of California. Seoul Bike Sharing Demand, 2020. URL https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand.

[5] Bojan Perovic, Dardan Klimenta, Miroljub Jevtic, and Miloš Milovanović. The effect of different sky temperature models on the accuracy in the estimation of the performance of a photovoltaic module. 59:78–82, 11 2019.

[6] Seoul Open Data Plaza, 2023. URL https://data.seoul.go.kr/index.do.

[7] Jan Philip Richter. Github - Seoul Bike Sharing Demand, 2023. URL https://github.com/janphiliprichter/SeoulBikeSharing.