Università degli Studi di Padova

SEOUL BIKE SHARING

Analysis on demand estimation

Brocco Mattia - 2044714
Magliani Jacopo - 2040912

INTRODUCTION

OBJECTIVES

- □ Discover patterns in the **bike sharing** system of Seoul through data exploration and modelling
- □ Make use of the available features to derive a **regression model** to estimate the count of rented bikes
- □ **Produce insights** on the shared public transport for further development

8760

the number of observations

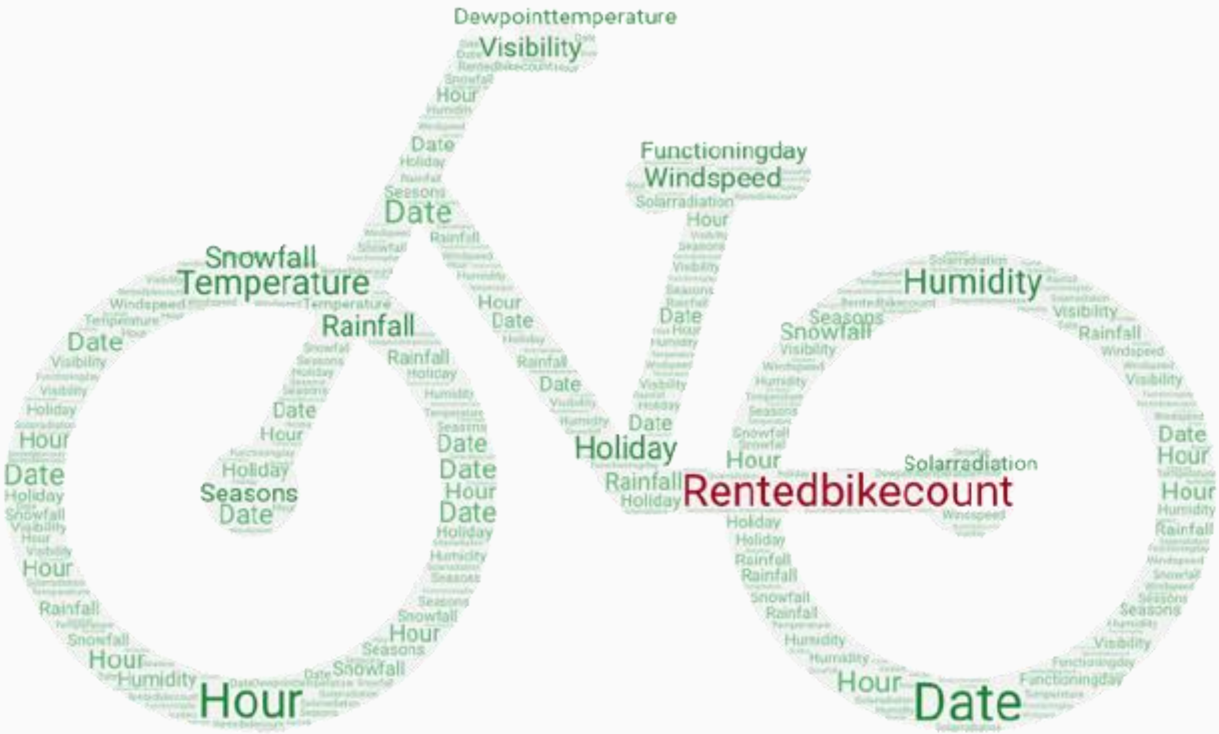
14

the number of variables

01/12 2017 | ... | 30/11 2018

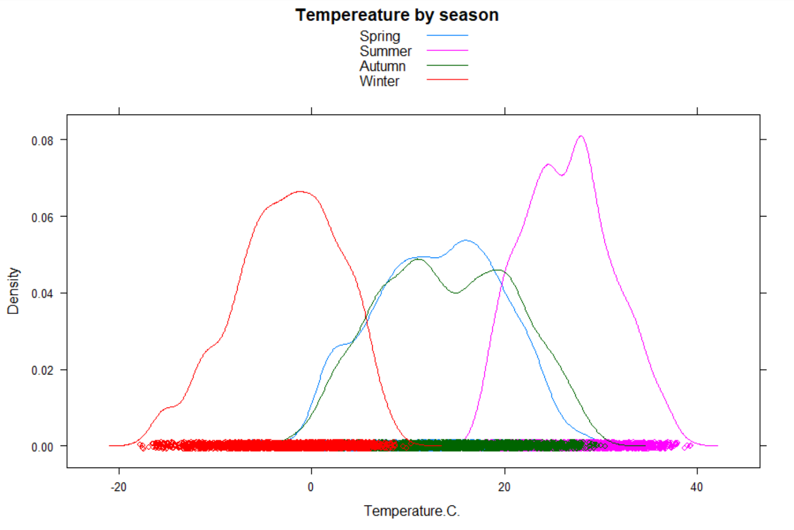
the time interval

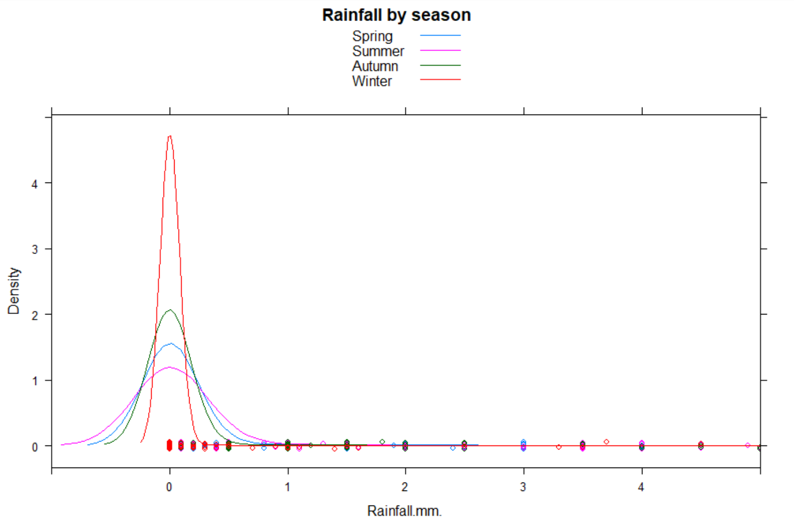
INTRODUCTION



DATA EXPLORATION

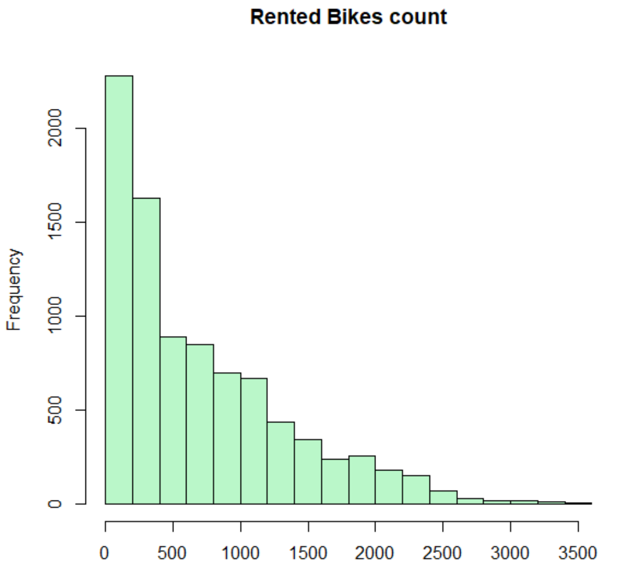
Atmospheric variables show trends not too far from what we expect from a continental climate throughout the different seasons (e.g., higher temperatures in Summer, lower rainfall in Winter).

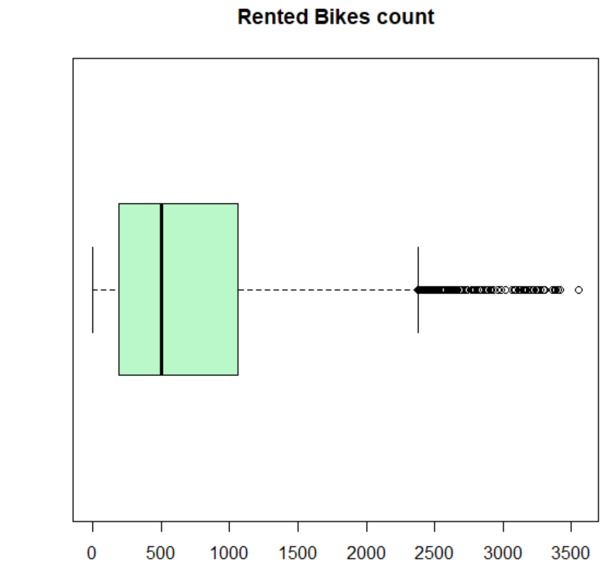




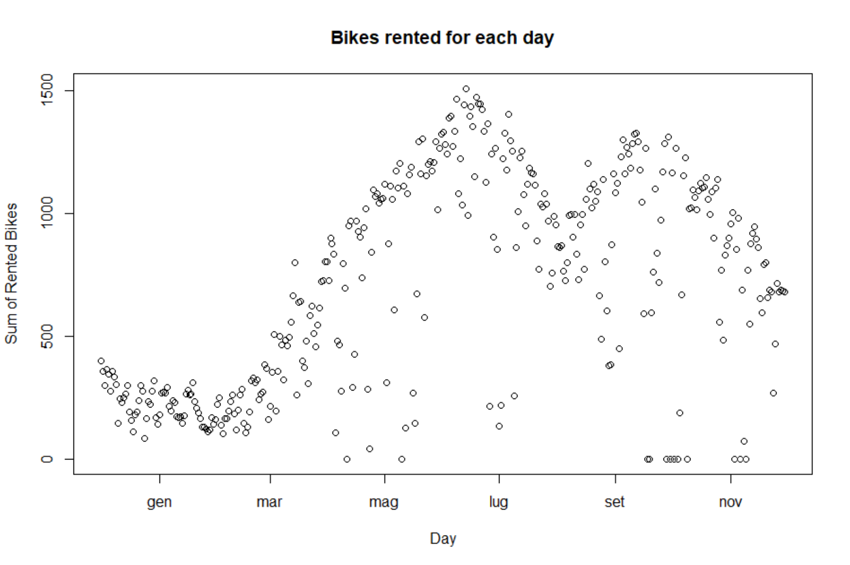
DATA EXPLORATION

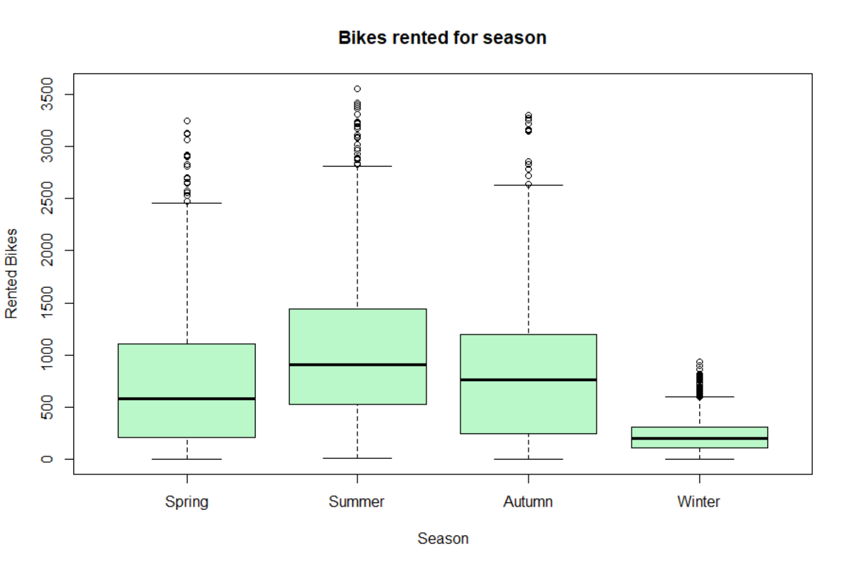
The target variable shows a left-skewed distribution, far from a normal distribution.





DATA EXPLORATION

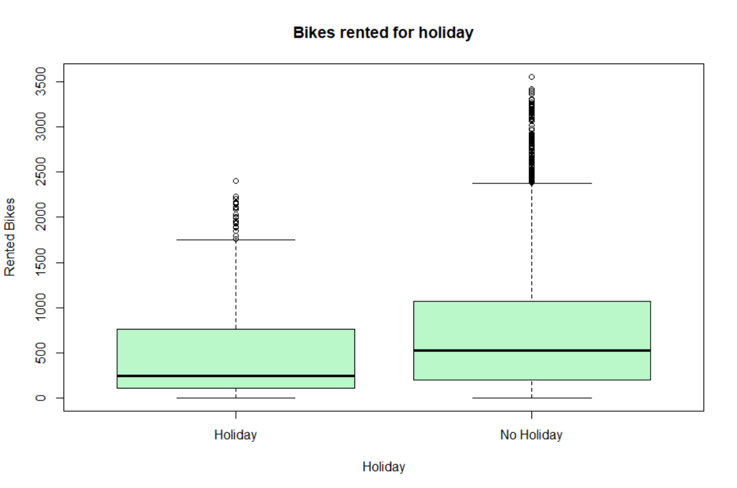


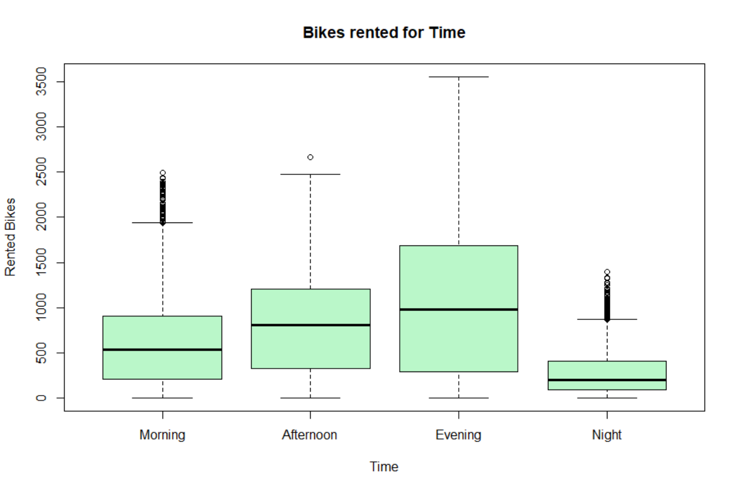


DATA EXPLORATION

- 18 holiday days
- Non-Holiday show an **average of rented bikes greater** than that of Holidays days. This may be also due to the small number of Holiday.
- The influence of a **specific holiday** can't be understood as the time interval considered is of 1 year.

Consideration on the target variable with respect to an additional **factor variable computed from the Hour** in a day.



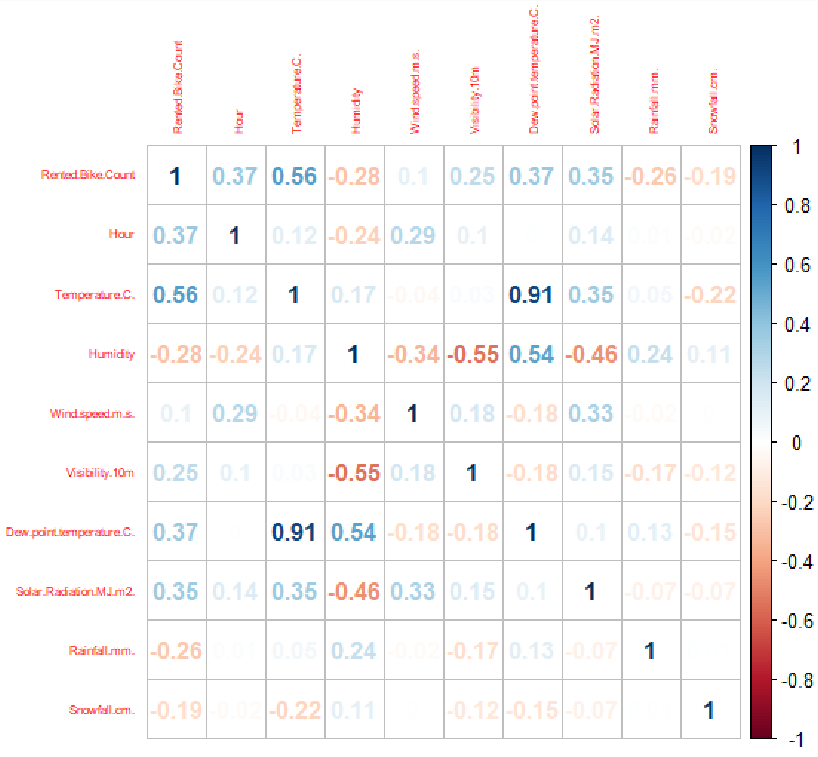


DATA PRE-PROCESSING

- Removed examples with **Functioning Day** flag **NO**
- Removed variables:
  Date, Time and Functioning Day
- Target variable: applied with log to make it more similar to normal distribution

DATA PRE-PROCESSING

There is a strong natural correlation between Temperature (°C) and Dew Point Temperature (°C)



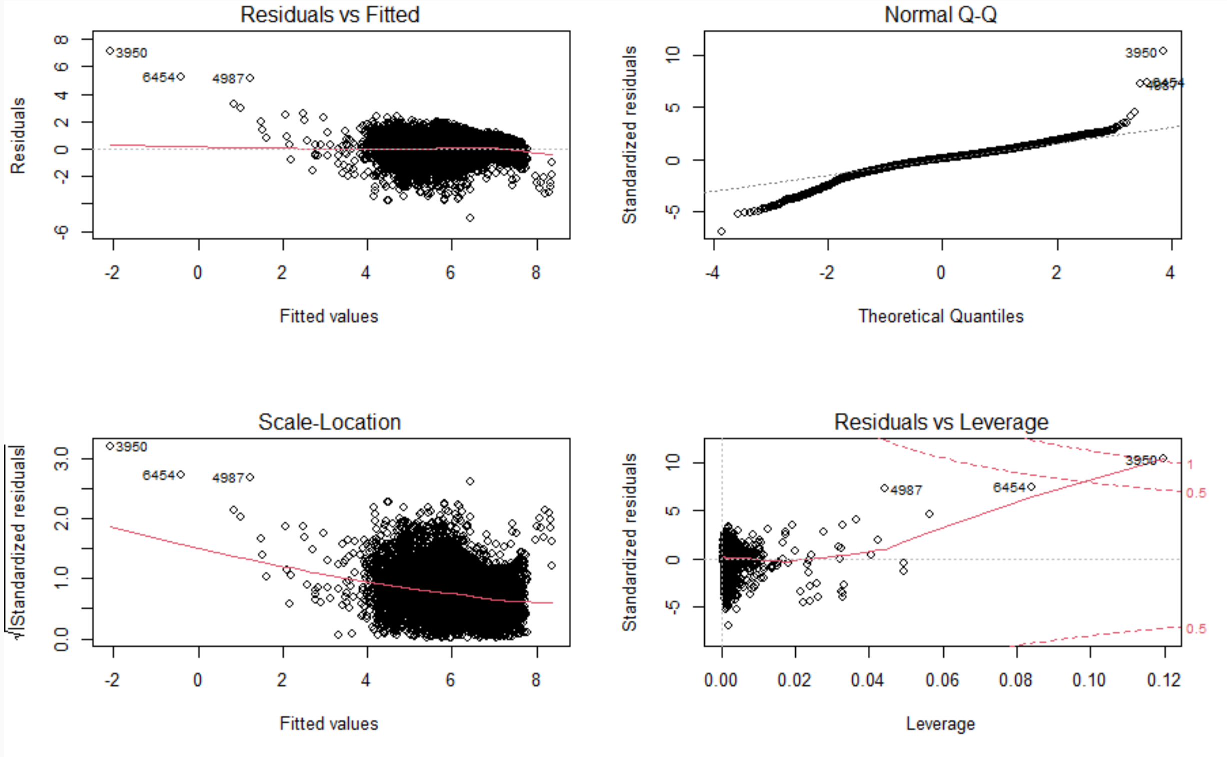
	Rented.Bike.Count	Hour	Temperature.C.	Humidity	Wind.speed.m.s.	Visibility.10m	Dew.point.temperature.C.	Solar.Radiation.MJ.m2.	Rainfall.mm.	Snowfall.cm.
Rented.Bike.Count	1	0.37	0.56	-0.28	0.1	0.25	0.37	0.35	-0.26	-0.19
Hour	0.37	1	0.12	-0.24	0.29	0.1	0	0.14	0.01	-0.02
Temperature.C.	0.56	0.12	1	0.17	-0.04	0.03	0.91	0.35	0.05	-0.22
Humidity	-0.28	-0.24	0.17	1	-0.34	-0.55	0.54	-0.46	0.24	0.11
Wind.speed.m.s.	0.1	0.29	-0.04	-0.34	1	0.18	-0.18	0.33	-0.02	0
Visibility.10m	0.25	0.1	0.03	-0.55	0.18	1	-0.18	0.15	-0.17	-0.12
Dew.point.temperature.C.	0.37	0	0.91	0.54	-0.18	-0.18	1	0.1	0.13	-0.15
Solar.Radiation.MJ.m2.	0.35	0.14	0.35	-0.46	0.33	0.15	0.1	1	-0.07	-0.07
Rainfall.mm.	-0.26	0.01	0.05	0.24	-0.02	-0.17	0.13	-0.07	1	0.01
Snowfall.cm.	-0.19	-0.02	-0.22	0.11	0	-0.12	-0.15	-0.07	0.01	1

REGRESSION MODELS – FULL MODEL

Adjusted R^2: **0.6058**

Variables whose estimate are **not significantly different from zero**:

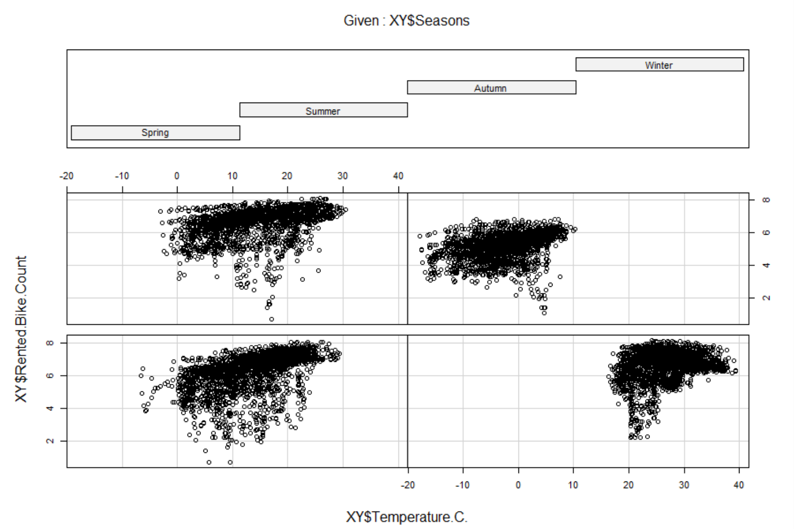
- Wind Speed
- Visibility
- Solar radiation
- Snowfall
- Season = Summer

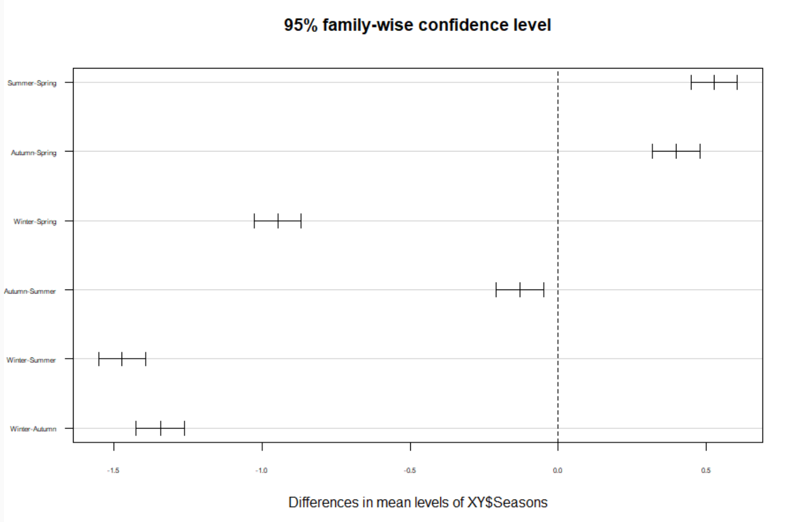


REGRESSION MODELS – FULL MODEL

ANOVA on levels of Seasons

1. Rejected H0 of Bartlett’s test on equal variances
2. Performed Fisher’s test anyway and rejected the null hypothesis of equal means between levels
3. Performed Tukey’s test on pairwise differences in means and rejected all null hypothesis on equality





CHANGES IN THE MODEL

GVIF of Dew Point Temperature > 10 provides evidence to remove the variable

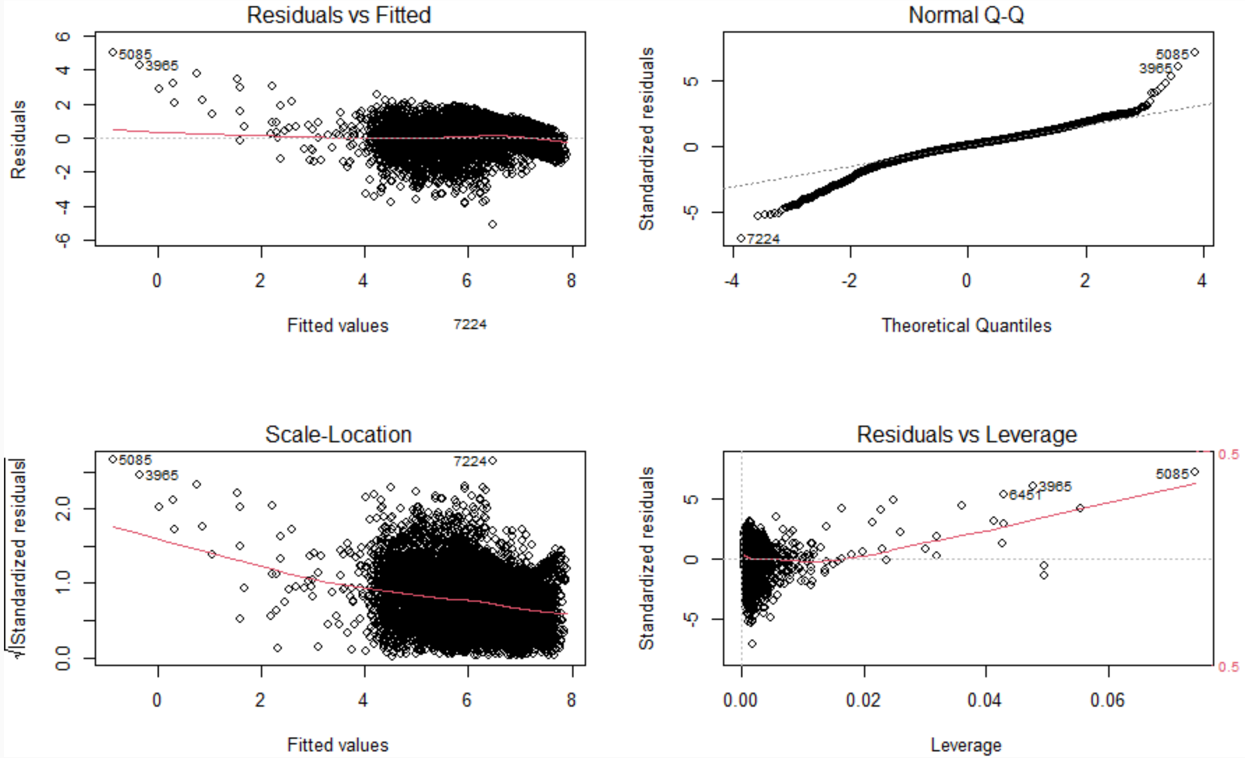
3 outliers removed

REGRESSION MODEL II

Adjusted R^2: **0.6125**

Collinearity problems **removed** (GVIF now all around 1)

ANOVA table shows **all regressors are significant**



CHANGES IN THE MODEL

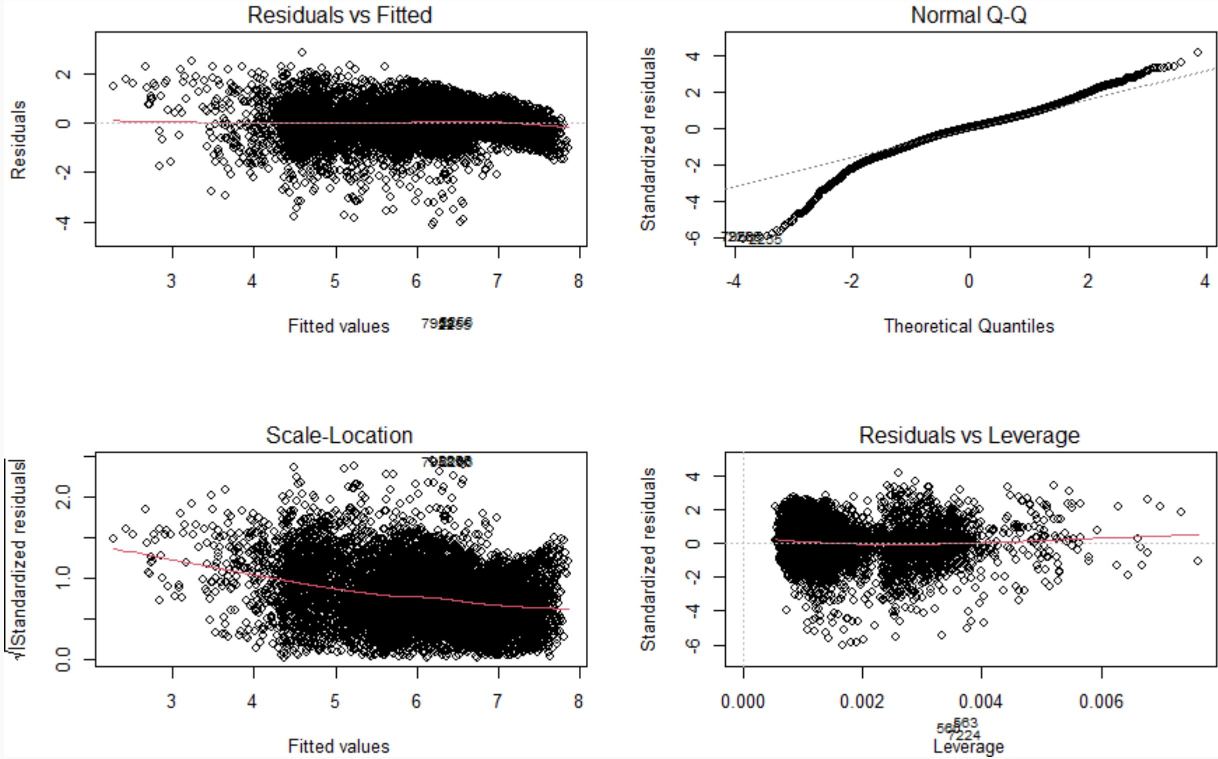
Binary factor variables for:

- Snowfall
- Rainfall

REGRESSION MODEL III

Adjusted R^2: **0.6531**

No changes in collinearity and against the assumptions of the linear regression.



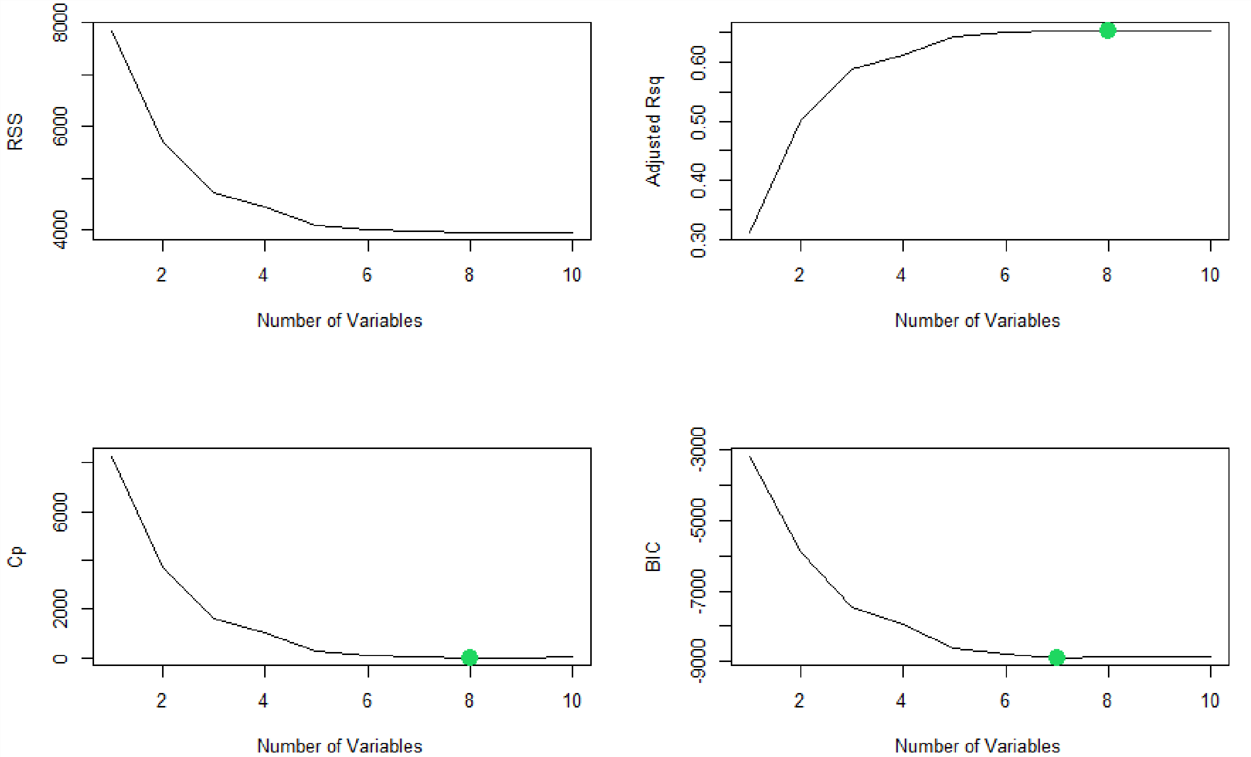
VARIABLES SELECTION

Are all variables gathered necessary?

- **Best Subset Selection**
- Backward step-wise selection
- Forward step-wise selection

Turns out the explained variance does not decrease when the following are **removed**:

- Visibility
- Wind Speed
- Snowfall



VARIABLES SELECTION

METHODS

Best Subset, Forward step-wise and Backward step-wise show **identical behavior**

CROSS-VALIDATION

We considered the output of Best Subset Selection, and confirmed it through **10-fold cross-validation**

TEST RESULTS

ANOVA (Chi-squared) test on best selection against full model with **p-value 0.9693**

ADJ. R^2

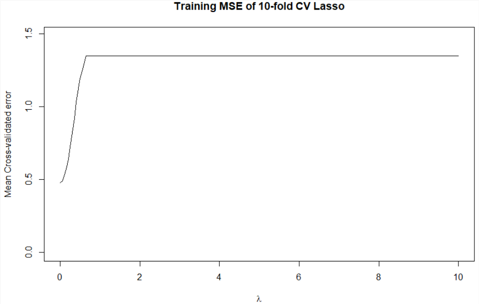
Same of the full model,

Adj. R^2 = 0.6532

SHRINKAGE METHODS

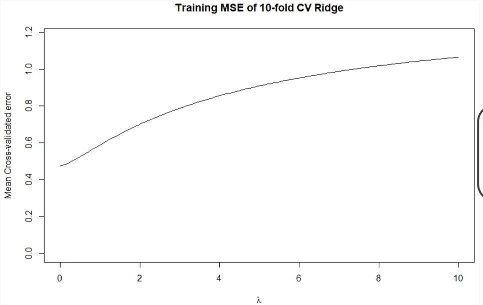
LASSO

Lambda: **0.1**



Training MSE	**0.5095**
Test MSE	**0.5164**
R^2	**0.6182**
Adj. R^2	**0.6177**

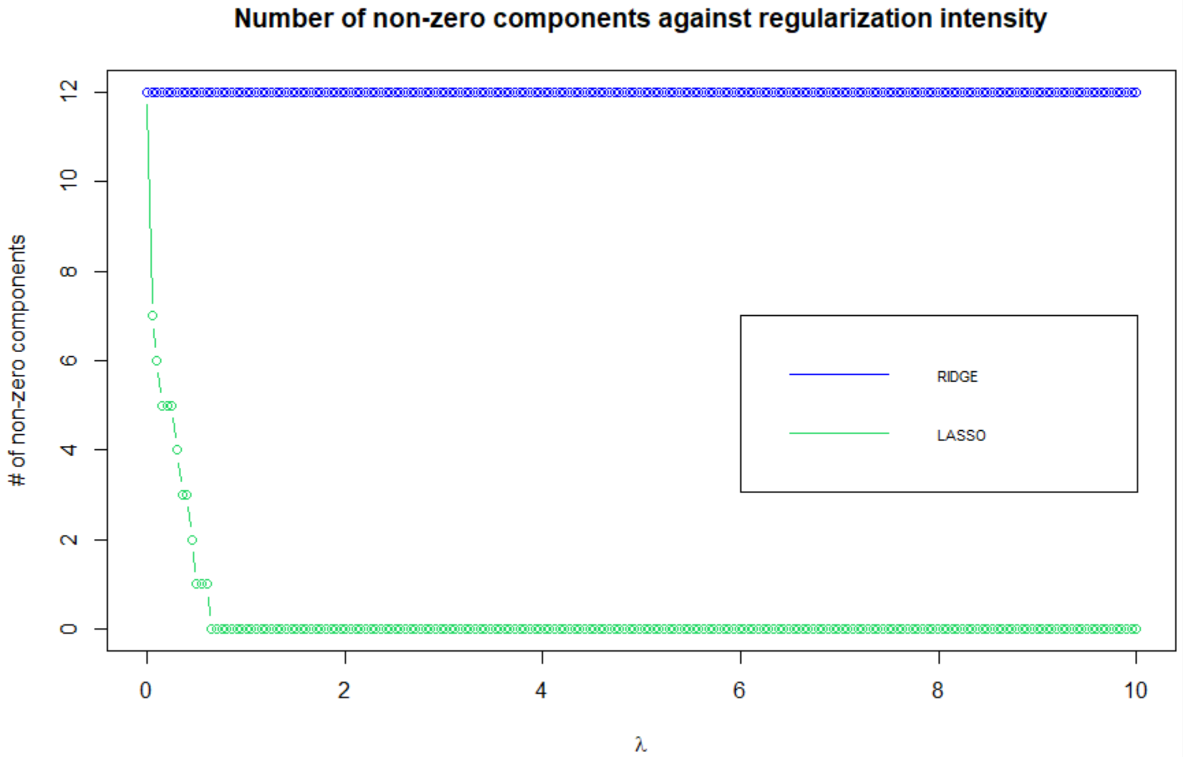
RIDGE



Lambda: **0.25**

Training MSE	**0.4793**
Test MSE	**0.4887**
R^2	**0.6387**
Adj. R^2	**0.6382**

SHRINKAGE METHODS – DETAIL

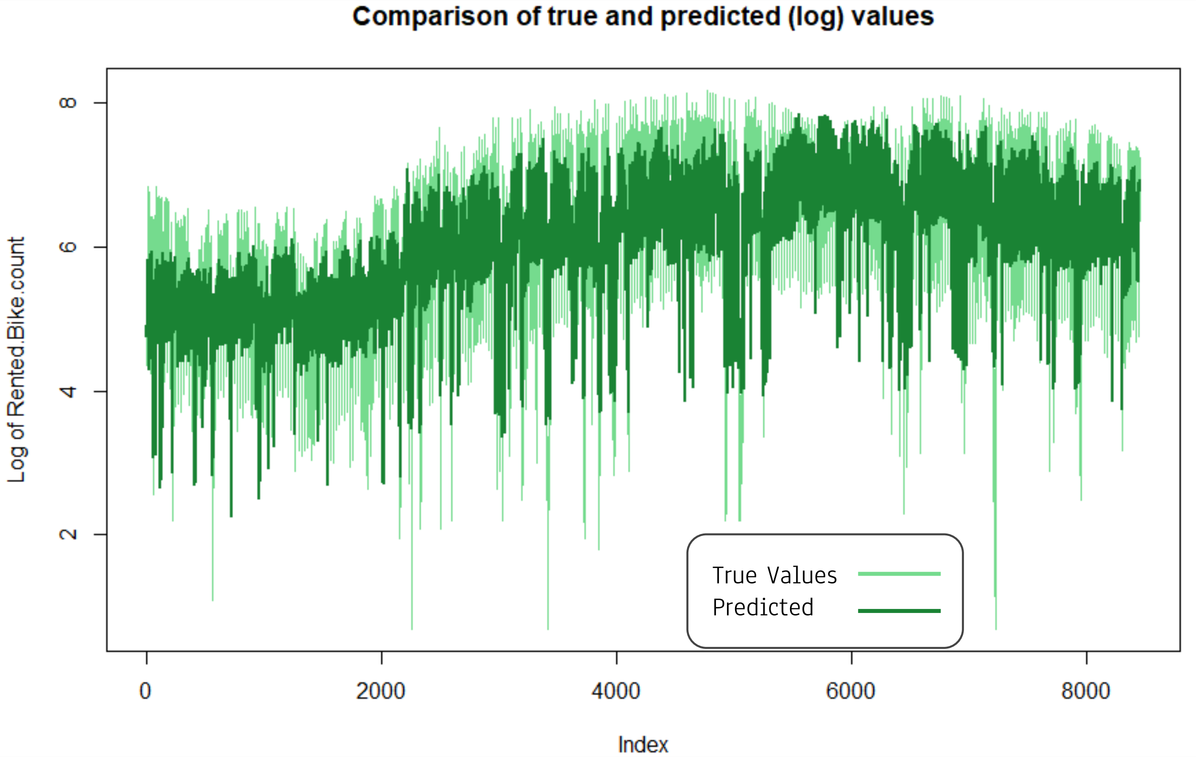


MODEL SELECTION

Given results on Variable Selection and Shrinkage, the **combined use** of the Best Subset variables and Ridge was tried, but with no improvement.

In this case **lambda** was even **lower** due to a simpler model.

FINAL CHOICE
Non-regularized model with only a subset of variables that estimates the log of the response variable.



PERFORMANCE RECAP

0.6058	(1) **All variables**, only «functioning days», **log** of Rented Bike Count
0.6125	(2) Outliers **removed**, excluded Dew Point Temp. due to **collinearity**
0.6531	(3) Transformation of **Snowfall** & **Rainfall** in binary factors
0.6532	(4) **Best Subset Selection**: excluded three variables
0.6177	(5) **LASSO** shrinkage from model described in **(3)**
0.6382	(6) **RIDGE** shrinkage from model described in **(3)**
0.6497	(7) **RIDGE** shrinkage from model described in **(4)**

CONCLUSION

VARIABLES

The **exclusion** of Dew Point Temperature, Wind Speed, Visibility, Snowfall and the use of a binary factor for Rainfall.

MODELS

The best performing model comes from the **selection of the best subset** of variables **without** the use of any **regularization**.

FURTHER WORK

Atmospheric variables may not be enough. Variables related to **road traffic**, presence of **events**, data from general **public transportation**, etc. may help in estimating Rented Bike Count.

Q&A

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Storyset**