Cole Gulledge, Maddox Johnston
Team Name: Regression Obsession
https://github.com/colegulledge/Bike_Share_Analysis_UCI_data

An Analysis of Bike Rentals by User Type through Ridge, Lasso, and Group Lasso Penalization Techniques

Abstract

This project seeks to create a predictive model for the number of hourly bike rentals from a Washington D.C. bikeshare using the Lasso and Ridge Regression. By further dividing the rentals into casual and registered user rentals, we observed differing levels of predictability for the features present in the models. Later, the Group Lasso model was employed to eliminate predetermined groups of multicollinear variables such as Days of the Week, Months, and Hourly Categorical Descriptions of Weather Conditions.

Introduction

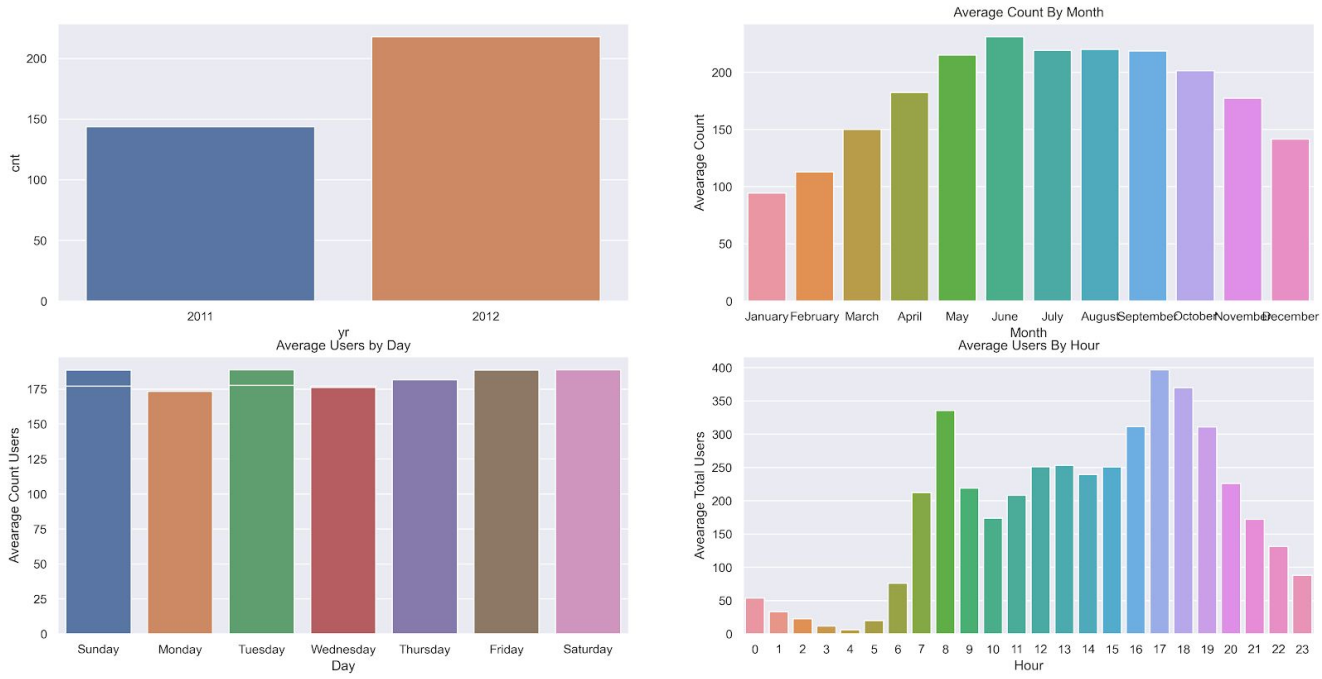
In urban cities across the world, bike sharing systems have become a popular, cheaper alternative to ride sharing apps and other conventional means of transportation. An individual can rent a bike from multiple different locations across town, and return it to an entirely different location, with no human interaction necessary. Bikeshares are notably a distributed system, thus posing an interesting possibility for research and widespread applicability.

Data Exploration

The dataset was found in the University of Irving Machine Learning Repository, and was donated by Hadi Fanaee-T. The dataset contains the number of hourly bike rentals from a Washington D.C. bikeshare over the course of two years (January 1st, 2011-Decemeber 31st, 2012) with numerous features and a grain of one hour. The dataset is found to contain no missing values *(see Figure C in the Appendix).* Throughout 17,380 rows spanning 731 days of bike rentals, the dataset notes the specific hour (0:00-23:00), day of the week (0-6), month (1-12) and a binary feature noting whether it is a weekend or not, and whether it is a holiday or not. In addition to these features, the author of the dataset includes corresponding weather features relating to each hour. The weather features present are: Temperature, Apparent ("feels like") Temperature, Humidity, Windspeed, and a brief categorical description of the conditions, i.e, Clear, Mist, Light Rain/Snow, or Heavy Rain/Snow. Temperature and Apparent Temperature are both

normalized in Celsius and derived as such: $(t_i - t_{,min})/(t_{max} - t_{min})$. The categorical variables above were dummy encoded, whereas the numerical weather features were left alone

There are three options for endogenous variables: Casual Rentals, Registered Rentals, and Count, or the total number of rentals- Casual and Registered- per hour. A Casual Rental is defined as a rider who is a guest and uses the service for a single ride, or up to 72 hours. Correspondingly, a Registered Rental is a rental from a registered rider possessing an annual membership which allows unlimited rides for up to a year. Casual Rentals are often tourists, while Registered Rentals are more likely to be locals that live within the greater area using the service as part of their daily commute.

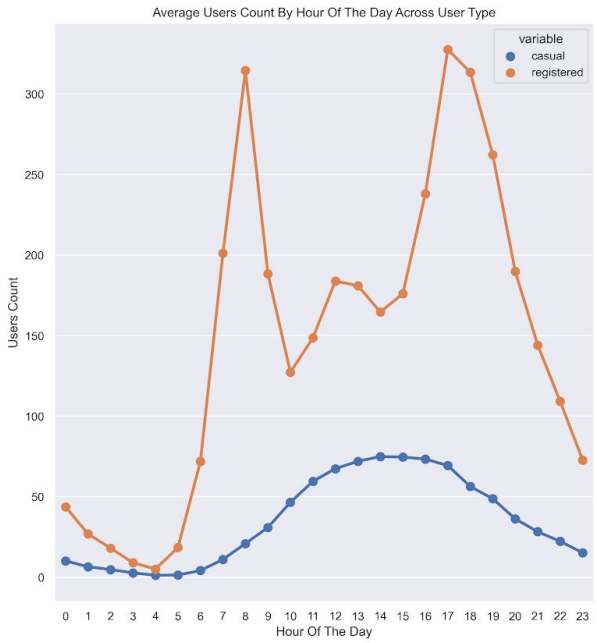


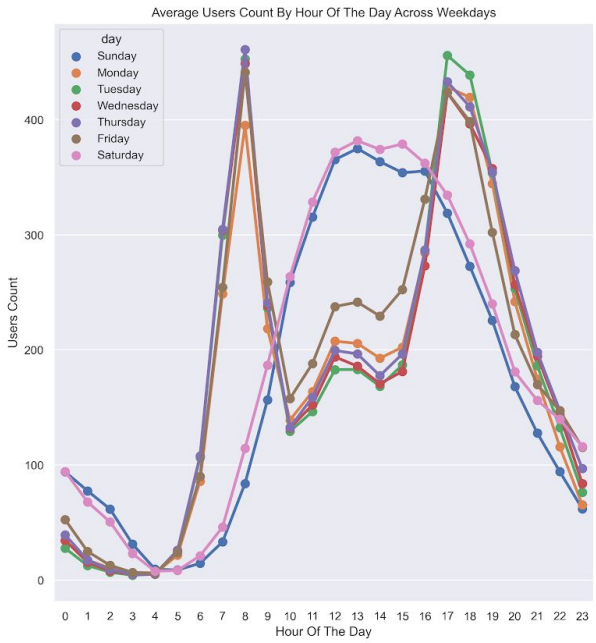
From upper left: Figure 1 denotes the average number of daily bikeshare users in 2011 and 2012. Figure 2 portrays the average number of daily bikeshare users per month. Figure 3 shows the average number of bikeshare users across the days of the week. Figure 4 denotes the average number of bike rentals, by hour (Here, hour 0 is 0:00).

By exploring bike rental counts over different time intervals, noteworthy trendys begin to appear . Figure 1 shows that daily bike rentals increased by nearly 35% from 2011 to 2012. Also, Figure 2 (*upper right*) shows that the Bike Share is seasonal, with individuals preferring to use the service in warmer months. Similarly, Figure 3 suggests that bikes rentals are slightly more popular over the weekend, as opposed to the weekdays. Lastly, Figure 4 (*bottom right*) shows

there is a massive spike in hourly bike rentals during the rush hour periods of 7:00-9:00 and 16:00-19:00, perhaps indicative of commuters using the service.

We will now further divide our daily bike rentals into the subcategories of Casual Rentals Registered Rentals to further elucidate hourly trends.





From left: Figure 5 shows the average casual (blue) vs. registered (orange) bike rentals, hourly. Figure 6 shows average hourly total bike rentals, colored by day of the week

Figure 5 separates the types of members by their membership type, allowing us to notice carried out by Casual Users that are less erratic throughout the day when compared to registered rentals, which appear to be heavily weighted toward popular arrival/departure times. Figure 6 shows the aggregate number of total hourly users throughout the day, coloreby the day of the week. We notice bike rentals on Saturday and Sunday seem to be most popular in the middle of the day (10:00-16:00), while Weekday rentals are most popular during “rush hours” . This seems to suggest that weekday rentals are heavily influenced by Casual Users, in contrast to the Registered User dominated weekday rental trend.

After noticing these trends, data points outside of three standard deviations of the Count variable were removed, and the data is prepared for analysis.

Analysis

The **Lasso** ("Least Absolute Shrinkage and Selection Operator") regression is often employed to create a sparse predictive model:

$$\min R(\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

Lasso Penalty Formula

The Lasso model is marked by the presence of the L1 norm penalty term, where the parameter λ is tuned to control the level of sparsity applied on the model. Here, variables with little contribution are disregarded entirely.

The **Ridge** regression uses the L2 norm to shrink non influential coefficients toward 0, while still considering them in the model:

$$\min R(\beta) + \lambda \sum_{i=1}^{p} (\beta_i)^2$$

Ridge Penalty Formula

Here λ is tuned to control the shrinkage of the model applied by the penalty term where large λ values are associated with more shrinkage.

Model	Lambda	Target Variable	Train R^2	Train RMSE	Test R^2	Test RMSE
		Total Users	0.25571	0.85903	0.24597	0.87938
LASSO	0.05	Casual Users	0.39944	0.77430	0.30612	0.83486
		Registered Users	0.17845	0.90060	0.16640	0.92990
		Total Users	0.29434	0.83643	0.27557	0.86194
RIDGE	0.05	Casual Users	0.44443	0.74470	0.35001	0.80803
		Registered Users	0.21870	0.87830	0.17362	0.92594
		Total Users	0.28806	0.84010	0.27259	0.86372
LASSO	0.01	Casual Users	0.44121	0.74690	0.34297	0.81239
		Registered Users	0.21140	0.88230	0.17664	0.92425
		Total Users	0.29434	0.83643	0.27556	0.86195
RIDGE	0.01	Casual Users	0.44440	0.74470	0.35000	0.80803
		Registered Users	0.21872	0.87833	0.17361	0.92590

Figure 7: This table displays the performance of the Lasso and Ridge regression with lambda values of 0.05, and 0.01 using Total Users, Casual Users, and Registered Users as endogenous variables.

Figure 7 presents a description of the performance of the Lasso and Ridge Regression with varying levels of shrinkage/sparsity. We notice that the Ridge regression outperforms the Lasso for predicting casual user rentals, with the caveat of a less sparse model: the Lasso regression tuned to a lambda value of 0.05 only considers 7 out of our 25 original features (Humidity, Apparent Temperature, whether or not it is May, July, and October, and whether or not it is Saturday or Sunday) [*see Figure 8, below*]. Coefficient visualizations of all models can be found in the Appendix.

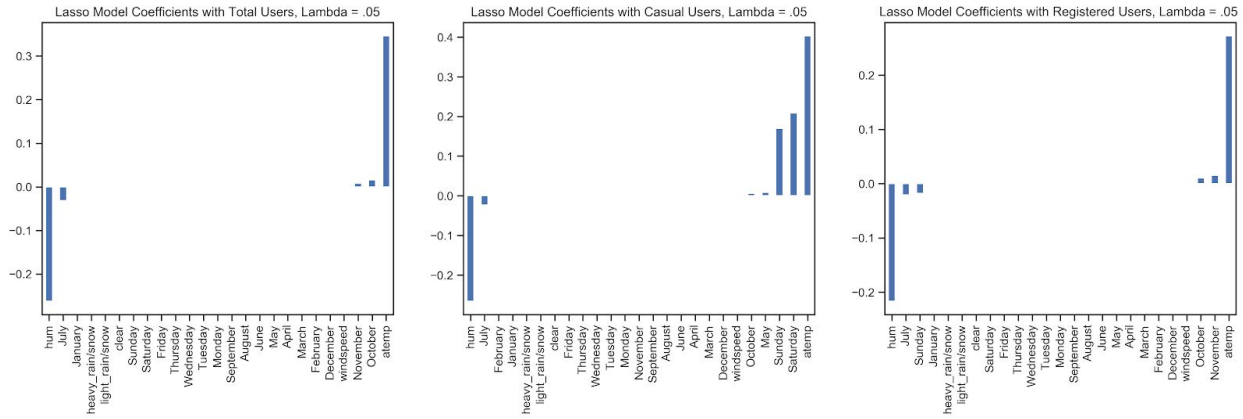


Figure 8: These graphs show the performance of the Lasso regression at a lambda value of 0.05 using Total Users, Casual Users, and Registered Users as endogenous variables.

This begets the natural question of if there is some way to control the consideration of these grouped features such as day of the week and month to create an even sparser model.

Group Lasso

The Group Lasso, proposed by Yuan and Li in 2006, is an extension of the Lasso regression which allows predetermined groups of multicollinear features to be considered (or not considered) as a group:

$$\min R(\beta) + \lambda \sum_{l=1}^{K} \sqrt{p_l} \left\| \beta^{(l)} \right\|_2$$

Group Lasso Penalty Formula

where p_l is the size of the l-th group. The Group Lasso is relevant for our problem since the Lasso regression tended to select some individual features within a group of highly correlated features to consider, while eliminating others. The Group Lasso allows our model to entirely disregard these multicollinear features while maintaining predictive applicability.

Unfortunately, scikit- learn has yet to adopt the Group Lasso penalization method within their library, so we were forced to utilize a third party library known as "asgl". Within its documentation, asgl is defined as "a Python package that solves several regression related models for simultaneous variable selection and prediction, in low and high dimensional frameworks". The Group Lasso class within this library does not yet have the ability to compute error metrics, so we elect to analyze its ability to eliminate pre-determined groups of features at different penalization parameters. We defined the groups as: Days of the Week, Months, and Categorical Descriptions of Weather Conditions, with each quantitative weather feature encompassing its own group. The results of these distinctly tuned models can be found in the first section of the Appendix [*Table A*].

Within the Total Users model, the Group Lasso with a penalization parameter of .05 was able to eliminate the Days of the Week group from its list of features, while the Casual Users model elected to eliminate the Categorical Description of Weather Conditions and Windspeed feature groups. These are highly interesting decisions from the Group Lasso, and can be explained by the Lasso model's varying coefficient weights, displayed in Figure 8 above. The Casual Users model places far more emphasis on which day of the week it was (we observe extremely popular Saturdays and Sundays [*see Figure 6*]), causing it to resist eliminating this influential feature group.

Future Work

Given more time, we would have liked to compute error metrics for the Group Lasso Models we employed. We would also possibly incorporate a Sparse Group Lasso model. These are relatively new models that have not yet been adopted by traditional python libraries; they may pose interesting results to analyze in the future.

References

Fanaee-T, Hadi, and Gama, Joao. (2013). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset]. Irvine, CA: University of California, School of Information and Computer Science.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.

Mendez-Civieta, A., Aguilera-Morillo, M.C. & Lillo, R.E. Adaptive sparse group LASSO in quantile regression. Adv Data Anal Classif (2020). https://doi.org/10.1007/s11634-020-00413-8

Yuan, Ming & Lin, Yi. (2006). Model Selection and Estimation in Regression With Grouped Variables. Journal of the Royal Statistical Society Series B. 68. 49-67. 10.1111/j.1467-9868.2005.00532.x.

Appendix

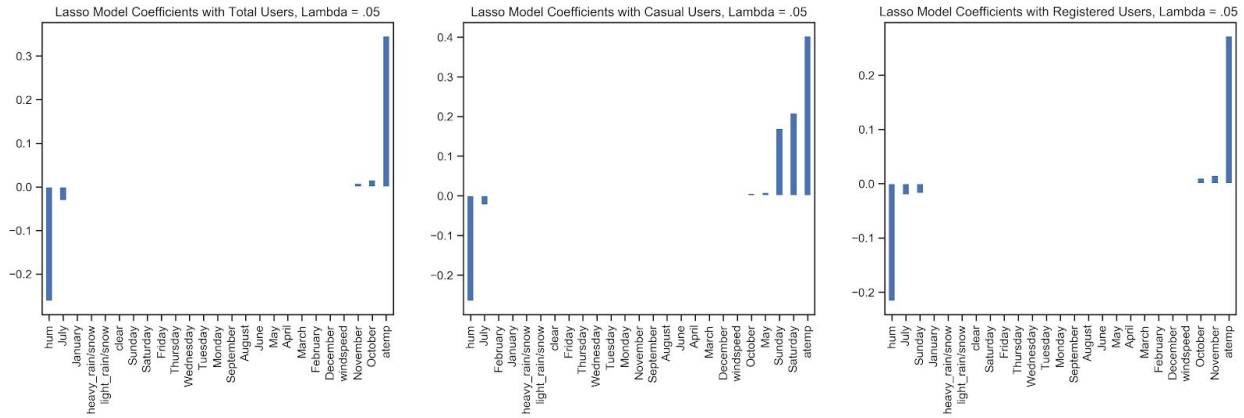
Table A

Group Lasso Coefficient Values at Varying Lambda Values

	Total User Model			Casual User Model			Registered User Model		
Feature	Lambda=0.001	Lambda=.01	Lambda=.05	Lambda=.001	Lambda=.01	Lambda=.05	Lambda=.001	Lambda=.01	Lambda=.0153
Intercept	-0.0038	-0.0038	-0.0037	-0.0020	-0.0019	-0.0013	-0.0040	-0.0041	0.0041
January	0.0445	0.0240	-0.0091	0.0538	0.0329	-0.0038	0.0350	0.0163	0.0080
Feb	0.0149	0.0012	-0.0147	0.0057	-0.0069	-0.0140	0.0160	0.0035	-0.0019
March	0.0087	0.0012	-0.0074	0.0425	0.0326	0.0059	-0.0046	-0.0102	-0.0123
April	-0.0025	-0.0032	-0.0020	0.0337	0.0304	0.0112	-0.0152	-0.0149	-0.0144
May	0.0061	0.0125	0.0147	0.0295	0.0340	0.0186	-0.0030	0.0037	0.0064
June	-0.0882	-0.0700	-0.0209	-0.0993	-0.0802	-0.0176	-0.0720	-0.0555	-0.0471
July	-0.1532	-0.1255	-0.0444	-0.1468	-0.1193	-0.0275	-0.1344	-0.1089	-0.0957
August	-0.0810	-0.0624	-0.0167	-0.0860	-0.0671	-0.0131	-0.0679	-0.0507	-0.0422
September	0.0079	0.0153	0.0170	0.0088	0.0160	0.0127	0.0064	0.0128	0.0151
October	0.0667	0.0621	0.0322	0.0555	0.0516	0.0188	0.0615	0.0566	0.0533
November	0.0921	0.0775	0.0302	0.0567	0.0443	0.0079	0.0922	0.0780	0.0703
Decemeber	0.0844	0.0675	0.0209	0.0459	0.0316	0.0006	0.0865	0.0695	0.0607
Monday	-0.0155	-0.0109	0.0000	-0.0504	-0.0483	-0.0378	-0.0014	-0.0007	-0.0004
Tuesday	-0.0135	-0.0091	0.0000	-0.0947	-0.0905	-0.0705	0.0169	0.0155	0.0145
Wednesday	-0.0075	-0.0052	0.0000	-0.0916	-0.0879	-0.0697	0.0231	0.0203	0.0187
Thursday	-0.0103	-0.0069	0.0000	-0.0983	-0.0942	-0.0738	0.0221	0.0200	0.0187
Friday	0.0142	0.0099	0.0000	-0.0322	-0.0319	-0.0273	0.0290	0.0248	0.0225
Saturday	0.0279	0.0195	0.0000	0.2001	0.1924	0.1527	-0.0365	-0.0326	-0.0302
Sunday	0.0042	0.0023	0.0000	0.1618	0.1554	0.1226	-0.0522	-0.0464	-0.0429
clear	-0.0200	-0.0170	-0.0051	-0.0232	-0.0171	0.0000	-0.0162	-0.0148	-0.0139
light_rain/snow	0.0264	0.0242	0.0075	0.0097	0.0108	0.0000	0.0289	0.0262	0.0243
heavy_rain/snow	-0.0115	-0.0097	-0.0033	0.0151	0.0116	0.0000	-0.0195	-0.0164	-0.0149
atemp	0.5599	0.5089	0.3831	0.6140	0.5636	0.4277	0.4638	0.4171	0.3941
humidity	-0.3172	-0.3166	-0.2948	-0.3295	-0.3262	-0.2898	-0.2695	-0.2690	-0.2678
windspeed	0.0348	0.0317	0.0158	0.0095	0.0084	0.0000	0.0391	0.0352	0.0328

Figure A

Coefficient Values of Lasso Models at Varying Lambda Values



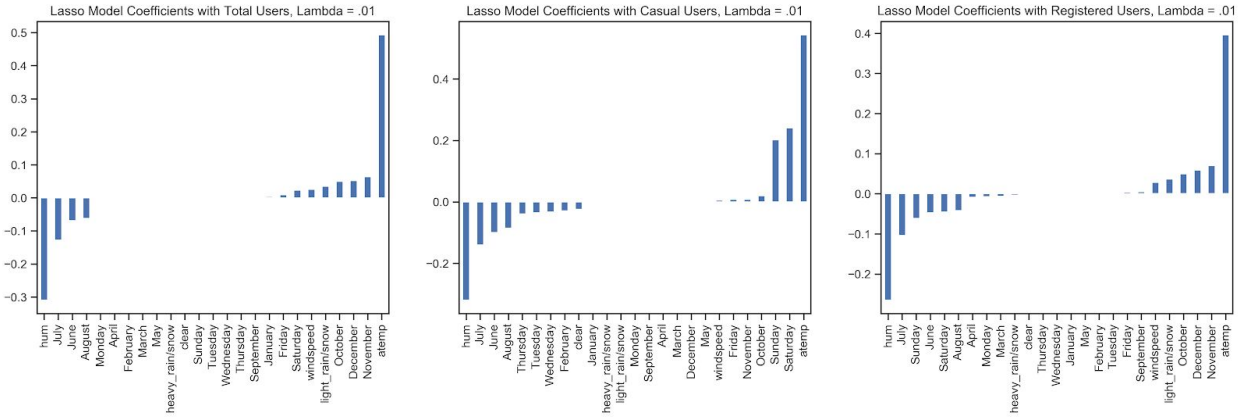


Figure B

Coefficient Values of Ridge Models at Varying Lambda Values

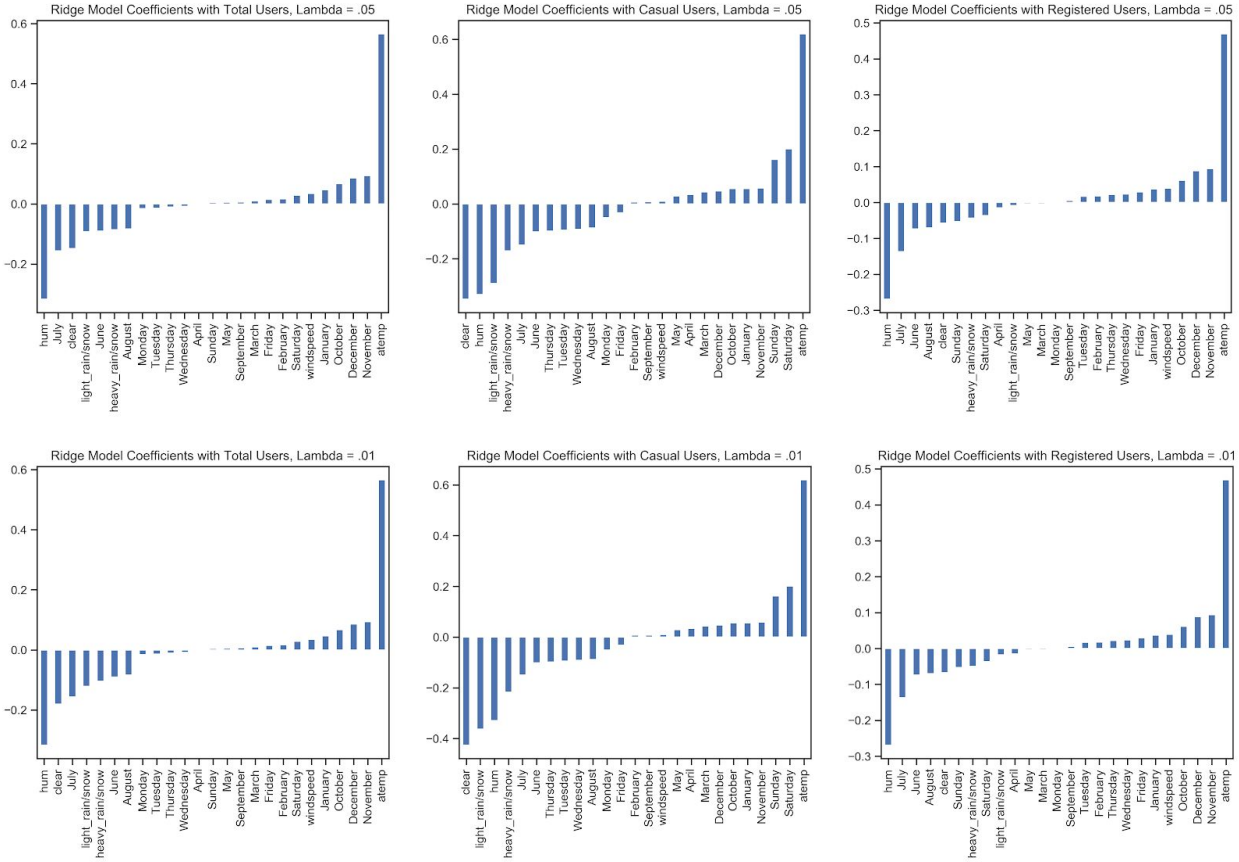
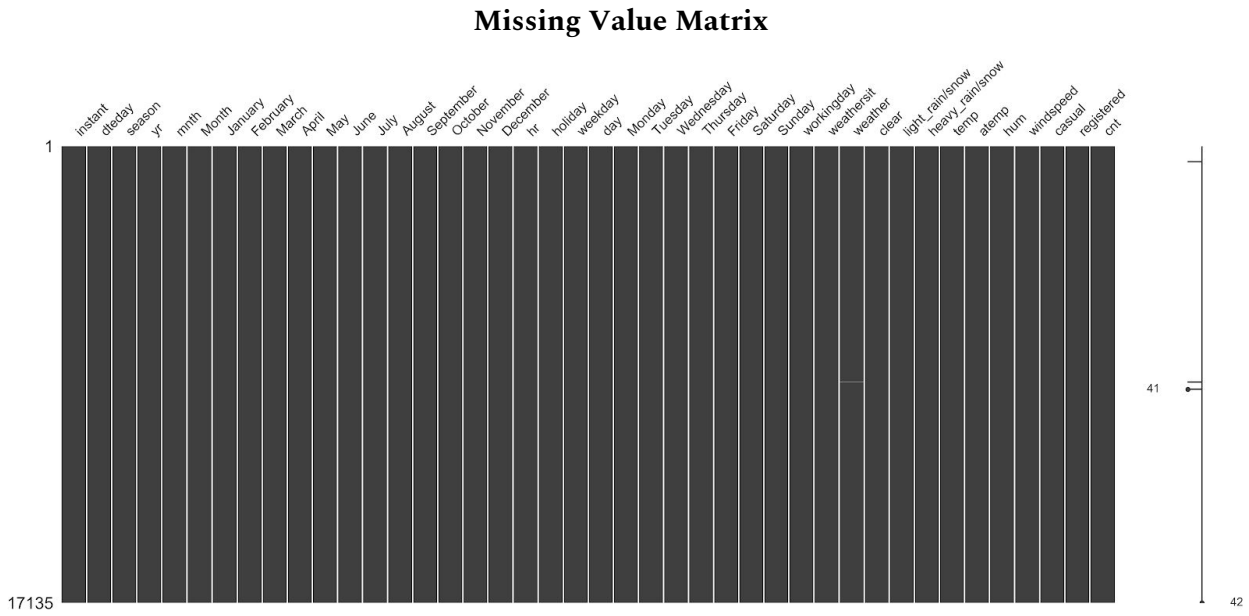


Figure C


Missing Value Matrix
instant
dteday
season
yr
mnth
Month
January
February
March
April
May
June
July
August
September
October
November
December
hr
holiday
weekday
day
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
workingday
weathersit
weather
clear
light_rain/snow
heavy_rain/snow
temp
atemp
hum
windspeed
casual
registered
cnt
1
17135
41
42