STAT 306 FINAL REPORT - A4

Group Members: Gina Choi, Harnoor Shinh, Jasper Law, Sohbat Sandhu

Prediction Model for Bike Rental Counts based on Environmental & Temporal Factors in Washington DC

INTRODUCTION

In the realm of urban transportation, bike sharing systems have been recognized globally as innovative solutions to mitigate traffic congestion, advocate for the usage of sustainable transportation, and improve overall levels of public health. As a growing alternative to motorized forms of transportation due to financial and environmental concerns, there is an ever-increasing demand to gain insights into the world of bike rental trends for organizations such as such cycle-share companies and urban planners.

Our objective is to use the Bike Sharing Dataset to build a robust prediction model for bike rental counts using explanatory variables such as temporal and environmental factors. This predictive model can be useful to provide actionable insights to optimize system operations for resource allocation and enhancing overall user experience. In the Central Bike Sharing (CBS) system used, when a successful rental occurs, the operative software collects basic data about the trip such as duration, start date, end date, etc, but it importantly also keeps a record of some of the key variables we are interested in: number of bikes rented and membership status.

As for the weather data, the authors of the dataset found a source that provides the Hourly Historical Weather Reports. They extracted some attributes such as weather temperature, apparent temperature, wind speed, wind gust, humidity, pressure, dew point and visibility for each hour from the period January 1, 2011 to December 31, 2012 for Washington, D.C., USA. Next, they mapped each hour in the bike rental time series with corresponding weather reports. Since there were missing weather reports for some hours, they mapped the closest report in the data for that hour.

Finally, they also extracted the Official holidays of Washington, D.C..and mapped them to the corresponding dates. Afterwards, the holiday dates were combined with weekends such that each day was classified into binary results of working day or non-working day. Additionally, they categorized each hour into four factors (weather grades): good, cloudy, bad and very bad, according to weather conditions provided in the weather data.

ANALYSIS

DATA INFORMATION

Variable	Data Type	Variable Description
instant	*Integer*	Record index for the bike rental
dteday	*Date*	Rental date

season	*Categorical*	Seasons labeled numerically with 1:winter, 2:spring, 3:summer, 4:fall
yr	*Binary*	Year of the rental: 0 indicates 2011 and 1 indicates 2012.
mnth	*Categorical*	Month of the rental labeled numerically with January corresponding to 1 and December to 12
holiday	*Binary*	Indicates whether the day of the rental is a holiday or not (extracted from District of Columbia holiday schedule)
weekday	*Categorical*	Day of the week
workingday	*Binary*	If day is neither weekend nor holiday variable is labeled 1, otherwise is 0
weathersit	*Categorical*	Type of weather encountered on the day of the rental: · 1: Clear, Few clouds, Partly cloudy · 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist · 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
temp	*Continuous*	Normalized temperature in Celsius
atemp	*Continuous*	Normalized feeling temperature in Celsius
hum	*Continuous*	Normalized humidity
windspeed	*Continuous*	Normalized wind speed (Original values are divided to 67 (max))
casual	*Integer*	Count of casual users for bike rental (SingleTrip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)
registered	*Integer*	Count of registered users for bike rental (Annual Member, 30-Day Member or Day Key Member)
cnt	*Integer*	Count of total rental bikes (casual users and registered users combined)

Table 1: Variables within our Dataset

PRELIMINARY DATA ANALYSIS

In our exploratory data analysis, we aimed to identify potential predictor variables from the BikeSharing Dataset that we could consider for further analysis and variable selection. We chose to exclude the record index, referred to as instant, as it served as a record identifier and held no relevance to our analysis. The variable representing the date, “dteday” was also removed for several reasons; with variables in the dataset indicating what day of the week it was, the month, the season, and whether it was a holiday, it appeared the date variable would not contribute significantly to our analysis. Furthermore, with only two years of data, utilizing the date as a predictor would not provide reliable insight due to potential atypical events and environmental phenomena occurring in this short timeframe.

Finally, we also chose to exclude the counts of casual and registered users from further analysis. Since the total count of bikes rented would be the sum of casual and registered users,

our model would be a perfect fitting model. In real-life applications, having the counts of casual and registered users would defeat the need for a model predicting the total count.

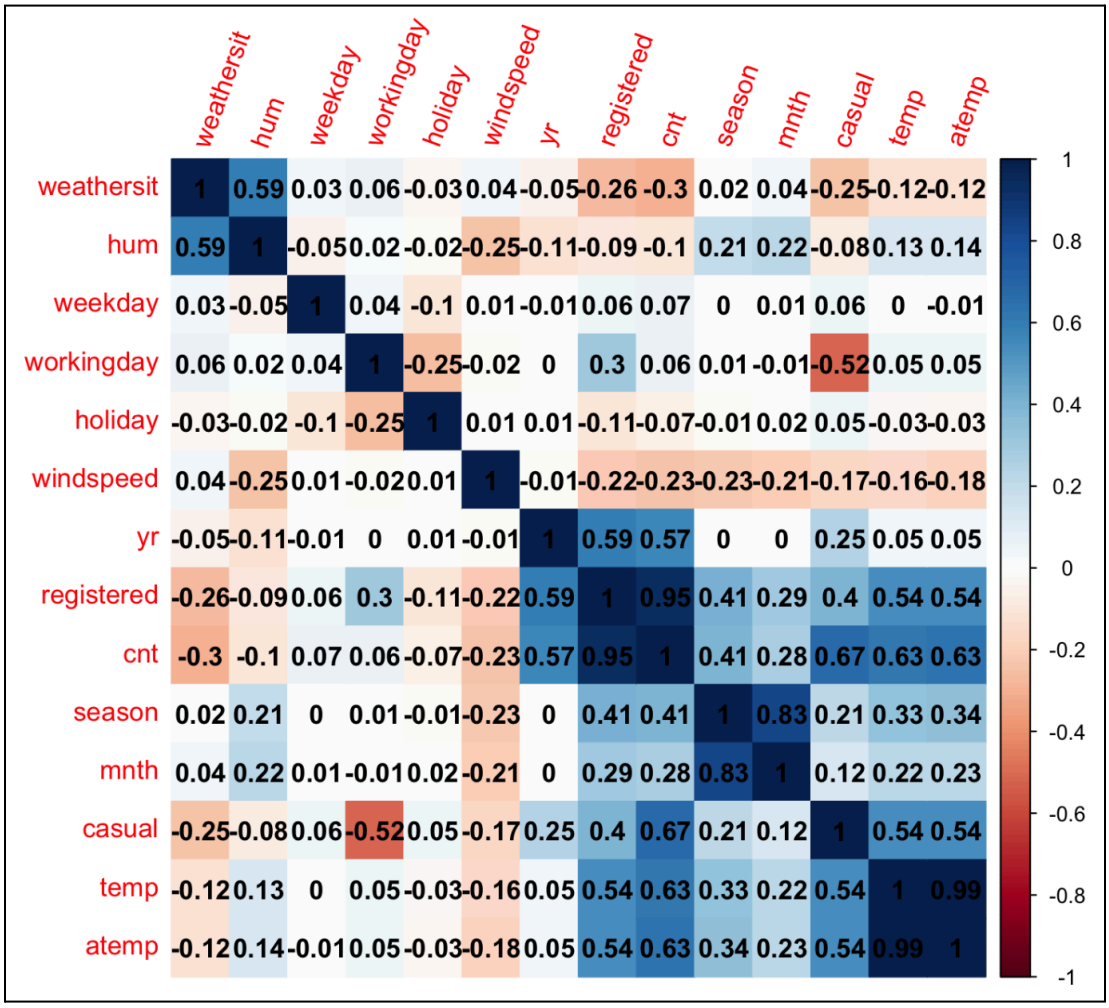
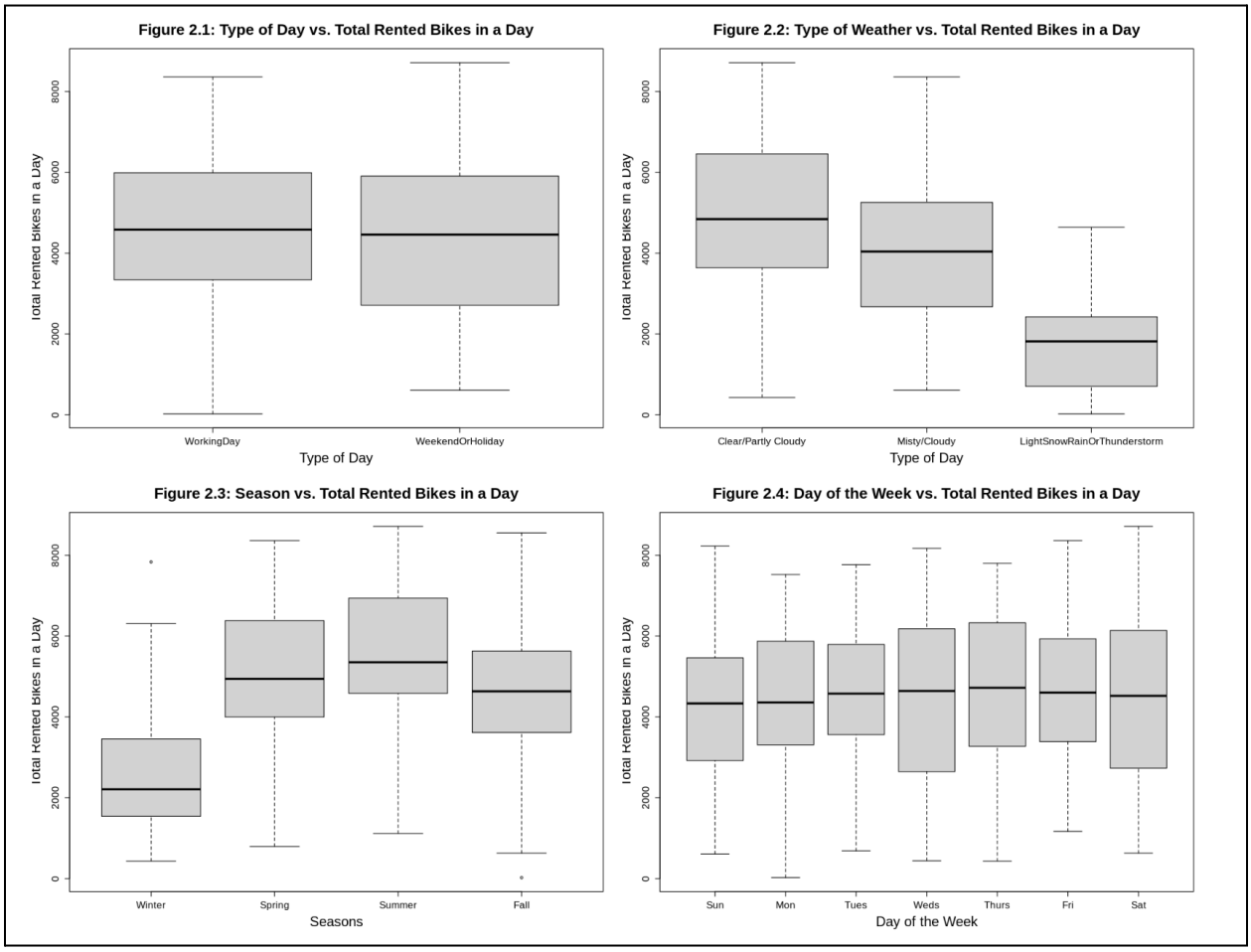


Figure 1. Heatmap to access Correlation

From the correlation matrix, we also observed a strong correlation of 0.83 between the “month” and “seasons” variable. Given all 12 months were included in the dataset, the consideration of month as a predictor would lead to the inclusion of 11 dummy variables. With running costs and model complexity in mind, we opted to utilize “seasons” instead to streamline the variable selection process and our model, bearing in mind the high correlation would likely lead to a similar model and explanatory power.

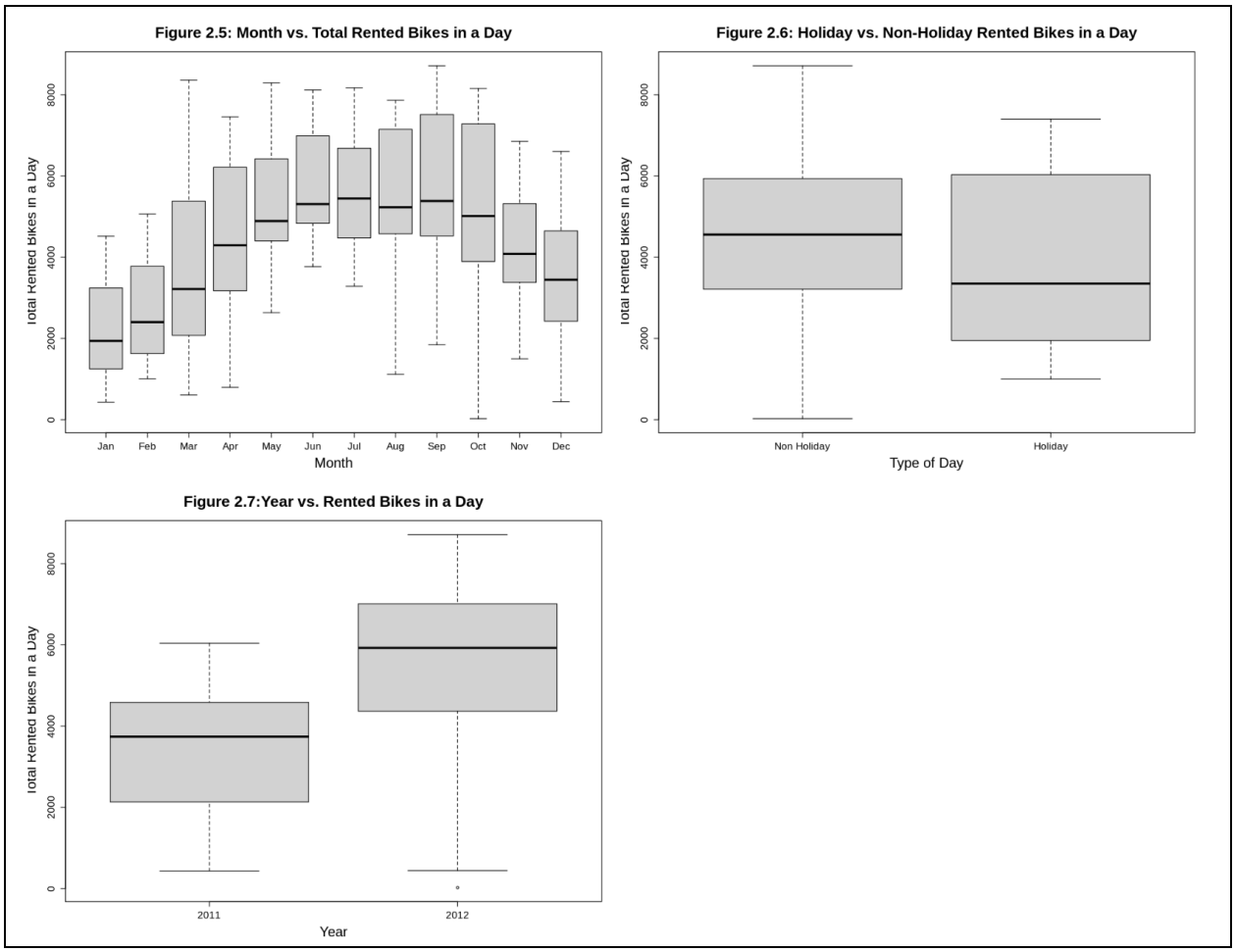
After examining potential linearity between the dataset variables, we found that the correlation matrix did not provide much insight into the relationship between bike rental count and variables such as weekday, workingday, weathersit, season, holiday, and yr. As this could be due to unaccounted non-linear relationships, we utilized side-by-side boxplots to visualize the data and investigate further. We plotted the categorical explanatory variables in the dataset against the total bike rental count per day, not accounting for multi-variable relationships and interactions.

Findings from this exploratory data visualization included discovering that weekends and holidays did not seem to have a notably different number of rentals compared to regular weekdays. The distributions of its respective box-plot, shown in Figure 2.1, appeared to have comparable centers and spreads. Similarly, days of the week did not seem to have a large effect on the number of users. All days shared similar medians and had light differences in distribution spread, shown in Figure 2.4.



Figures 2.1 to 2.4: Boxplots of categorical variables vs. Total Rental Bikes in a Day

Clear differences in distributions were seen in environmental factors such as the type of weather and the season it was. Naturally, we saw an increase of bike rentals in the summer and spring seasons (Figure 2.3), on clear and partly cloudy days (Figure 2.2) and a lower number of users in the winter, in rainy and snowy conditions. A notable increase in rentals from 2011 and 2012 was observed in Figure 2.7, suggesting a surge of popularity or expansion occurred. Holidays appeared to decrease the number of users compared to non-holidays, possibly as a result of people not requiring transportation to go to work. (Figure 2.6)



Figures 2.5 to 2.7: Boxplots of categorical variables vs. Total Rental Bikes in a Day

MULTICOLLINEARITY

Upon observing that the dataset contained variables for both normalized apparent and felt temperature, we opted to choose perceived temperature for variable selection over apparent temperature, subsequently excluding the “temp” variable in favour of the “atemp” variable. This was due to its greater relevance for cyclists, where the effects of speed, wind and humidity influence the suitability of temperature conditions for cycling. With a linear correlation of 0.99 from our matrix found in the correlation heatmap, including both “temp” and “atemp” would violate our independence assumption of linear regression modeling. In order to verify this we decided to perform VIF analysis with and without temperature included on our continuous variables and observed the following tables

VIF values	temp	atemp	hum	windspeed
Before	62.9699	63.6323	1.0782	1.1267
After	NA	1.0451	1.0764	1.0922

Table 2 . Variance Inflation Factors before and after removing highly correlated variable

Seeing this confirmed that we would need to remove either temp and that our other continuous variables do not exhibit much multicollinearity.

season	yr	holiday	workingday	weathersit	atemp	hum	windspeed	cnt
Winter	2011	NonHoliday	WeekendOrHoliday	Misty/Cloudy	0.363625	0.80583	0.1604460	985
Winter	2011	NonHoliday	WeekendOrHoliday	Mist /Cloudy	0.353739	0.69608	0.2485390	801
Winter	2011	NonHoliday	WorkingDay	Clear/PartlyCloudy	0.189405	0.43273	0.2483090	1349

Table 3. First 3 Rows of Transformed Dataset for Model Selection

MODEL SELECTION

As the purpose of this study is to do predictive analysis on the number of Bike rentals in Washington D.C., we will be implementing an **80/20 split** of the dataset into training and testing sets. The testing set will be later used to perform cross-validation using the **validation set approach** to evaluate the final model's prediction accuracy. The training set will be used to train the regression model. To identify the model with the highest predictive power from our full model, we will use the **Exhaustive** method in **Stepwise Selection Algorithm** to evaluate the linear model's goodness of fit. We will use model metrics such as Mallow's C_p, Adjusted R^2 and *BIC* (Bayesian Information Criterion). The final model will be checked for violations in linear model assumptions for further tuning.

From Table. 2, to conduct model selection, we will be choosing all the variables in the transformed data in our full model including the interaction terms:

- Normalized feeling temperature (atemp) and Normalized humidity (hum): Since humidity and temperature often contribute to how willing the average person is willing to go outside
- Normalized wind speed (windspeed) and Normalized feeling temperature (atemp): High amounts of wind can often make someone feel colder than what the temperature would normally dictate. In hotter months this would obviously encourage more outdoor activity, and in colder months would do the opposite
- Seasons (season) and Holiday (holiday): Many people often go out for certain holidays depending on the season which would likely contribute to bike rentals
- Normalized wind speed (windspeed) and Normalized humidity (hum): Since humidity depends on the amount of water vapor in the air, and wind speeds can decrease evaporation there is likely some interaction between these 2 variables
- Normalized feeling temperature (atemp) and Seasons (season): People's expectations of the temperature of a given season can often drive their plans for the day. For example it would make sense that a colder day in the summer would be less popular than a colder day in the Winter

After conducting the stepwise selection algorithm, we get the following metrics for:

Parameter (#)	BIC	Adj R2	Mallow's Cp	Parameter (#)	BIC	Adj R2	Mallow's Cp
1	-242.663	0.391845	1762.1961	**11**	-952.976	0.863246	13.2428
2	-572.925	0.684586	669.5865	**12**	-948.939	0.863560	13.0950
3	-648.019	0.730465	498.7593	**13**	-945.839	0.864122	12.0467
4	-703.308	0.760568	387.1230	**14**	-941.440	0.864336	12.2703
5	-762.372	0.788873	282.5841	**15**	-936.484	0.864403	13.0356
6	-824.853	0.815072	186.2398	**16**	-931.106	0.864356	14.2103
7	-868.603	0.831972	124.5693	**17**	-925.349	0.864208	15.7497
8	-907.841	0.845972	73.8159	**18**	-919.660	0.864078	17.2239
9	-944.001	0.857953	30.6862	**19**	-913.638	0.863858	19.0186
10	-955.935	0.862641	14.4617	**20**	-907.421	0.863586	21.0000

Table 4. Model Metrics observed for Exhaustive Stepwise Selection Algorithm

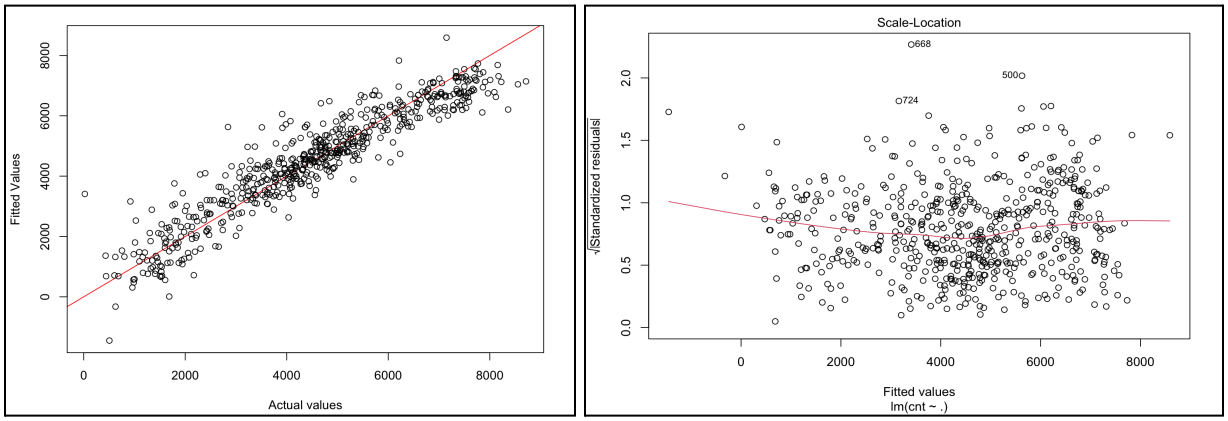
From Table 4, when observing the values for the C_p model metric we find the model 13 to be the one with the minimum C_p value with the associated Adjusted R^2 value of 0.864122 as this C_p value is quite close to the number of parameters in the model 13 (which includes 13 parameters). Model 13 is the best model that is able to strike a balance between model complexity and adequate model performance. Model 13 has the following explanatory variables stated in Figure 3. that explain 86.4122% of the variation in the number of Daily Bike rentals in Washington D.C.

```
Call:
lm(formula = cnt ~ ., data = training_subset)

Residuals:
   Min     1Q Median     3Q    Max
 -3215   -368     47    446   2183

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                             1262.2      238.0    5.30  1.7e-07 ***
yr2012                                  2030.0       63.1   32.15  < 2e-16 ***
`weathersitMisty/Cloudy`                -400.0       81.6   -4.90  1.3e-06 ***
`weathersitLightSnow/Rain/Thunderstorm` -2092.1      228.0   -9.18  < 2e-16 ***
seasonSpring                            1499.6      337.6    4.44  1.1e-05 ***
seasonSummer                            8057.5      580.0   13.89  < 2e-16 ***
seasonFall                              1487.1      110.0   13.52  < 2e-16 ***
holidayHoliday                          -458.2      227.5   -2.01    0.045 *
atemp                                   7389.5      494.8   14.94  < 2e-16 ***
hum                                    -1974.6      297.4   -6.64  8.3e-11 ***
windspeed                              -2677.8      443.7   -6.03  3.1e-09 ***
`seasonSummer:holidayHoliday`            908.3      473.8    1.92    0.056 .
`seasonSpring:atemp`                   -1303.1      744.9   -1.75    0.081 .
`seasonSummer:atemp`                   -11693.6      980.9  -11.92  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 707 on 497 degrees of freedom
Multiple R-squared:  0.868,     Adjusted R-squared:  0.864
F-statistic:  250 on 13 and 497 DF,  p-value: <2e-16
```

Figure 3. Summary of Reduced Linear Model

CHECK LINEAR MODEL ASSUMPTIONS

In order to test our model for violations of our linear model assumptions we used the following figures as violations of these assumptions can lead to misinterpretation of our results.



Figures 4.1-4.2 . Linearity: Fitted Vs. Actual Values; Homoscedasticity: Residuals Vs. Fitted Values

From our plot of fitted vs observed values, we can see that our assumptions of linearity holds true as the observed line very clearly follows a linear distribution

From our residuals vs fitted value plot we can also see that our homoscedasticity assumption holds as errors seem to have consistent variance as our fitted values increase

Our assumptions of independence of data and absence of multicollinearity are trivially true and therefore our model holds true with all of our assumptions required for linear modeling and we can begin using cross validation to assess how our model works on new data

CROSS VALIDATE WITH VALIDATION SET APPROACH

As there are no violations in the linear assumptions of the reduced fitted model in Figures 4.1-4.2 for our final fitted model in figure 3, we performed cross validation to assess RMSE of our final model in comparison to our initial full model.

Model	RMSE
Full Model	792.335
Final Reduced Model using exhaustive search	786.994

Table 5: RMSE for the Full model Vs. Final Reduced Model

From the data in table 5, we see that our final model has similar performance to our initial full model when applied to new data. This combined with the decreased complexity of our final model leads us to be confident in the validity of our final model in generalizing to new data.

CONCLUSIONS

In summary, our analysis delved into the relationship between various factors, primarily temperature, and the number of rented bikes. We found a significant correlation between apparent temperature and bike rental counts, highlighting apparent temperature as one of the

crucial determinants of rental behavior. Additionally, factors such as wind speed and humidity also exhibited strong correlation with counts of bikes rented.

“atemp” has the largest coefficient of our parameters - expected because the warmer the temperature it feels (up to a point), the more likely an average person will go cycling. “seasonSummer” also has a large coefficient - this is also in line with our predictions since more people tend to bike in the summer in hotter temperatures for exercise and personal fitness - the summer season also in general has fairly good weather for outdoor activities.

Some drawbacks about our model include its lack of interpretability - our model includes a decent number of parameters - which would take some time to explain in full detail.

An interesting note is that while you would expect (year-by-year) the data to vary in a cyclical fashion and be around the same for similar seasons each year, the data for 2012, on average, showed much higher bike counts for the entire duration of the year than 2011.

Overall, our findings display the complex relationships between a wide range of environmental factors and bike rental behaviour, offering valuable insights for urban mobility solutions. As we continue to refine our models and methodologies, we aim to contribute further to the advancement of sustainable transportation systems.

In conclusion, our comprehensive analysis has successfully addressed our research question by developing a robust prediction model for bike rental counts using temporal and environmental factors, displaying relatively good explanatory power.

REFERENCES

- Fanaee-T,Hadi. (2013). Bike Sharing. UCI Machine Learning Repository. https://doi.org/10.24432/C5W894.
- *Holiday schedules*. (n.d.). DCHR. https://dchr.dc.gov/page/holiday-schedules