EDA

Ricky Heinrich & Vimaljeet Singh

2023-03-10

Workflow for current and next step in the project

To determine which model would fit this data, we need to consider several factors such as the nature of the problem, the available data, and the type of output that we are trying to predict.

First, we need to define the problem and the goal of the model. In this case, we are trying to **predict the number of bikes rented per hour (cnt)**, given the various features such as weather, time of day, and other factors. We also want to know which predictors influence the number of bikes rented the most. This is a regression problem.

Second, we need to examine the available data and determine if there are any missing values, outliers, or other anomalies. If the data is incomplete, we may need to consider techniques such as imputation or data cleaning to address these issues.

Third, we need to select appropriate features for the model. Some features, such as season or weather, may have a strong correlation with the number of bikes rented, while others may not be as important. Feature selection can be done using techniques such as correlation analysis or principal component analysis.

Finally, we can select a model that is appropriate for the problem at hand. Some popular models for regression problems include linear regression, decision trees, random forests, and neural networks. We can use techniques such as cross-validation to evaluate different models and select the one that performs best on the data. From a multi linear regression, we may observe the p-value or use step-wise regression to determine which predictors are the most useful. In a random forest model, we may determine the importance of a predictor using the Gini Index.

Statistically Descriptive Analysis of the Dataset

Date related variables

The dataset contains 17 variables, where the first column is an index. Most of the data is numerical in nature, apart from the *dteday* column, which records the date in a date format. This date is separated further in year, month, and hour columns, which are factored into discrete numeric variables, similarly to the *weekday* variable, where Sunday is 0. The *year*, *holiday*, and *workinday* variables are boolean variables. For the *year*, a '0' represents 2011, and a '1' 2012.

We are expecting that the counts of observations for each year, weekday and hour are uniform across categories, as each should have the same number of hours. As shown in figure 1, this was not quite the case. We suppose that there is a dip in 'observations' for the hours of 2,3,4,5 due to some days containing no rides during that hour. The count of hours with rides is slightly lower in 2011 than 2012, and we see a small concave curve with Sunday and Saturday on both ends and where there minimum count is on Tuesday.

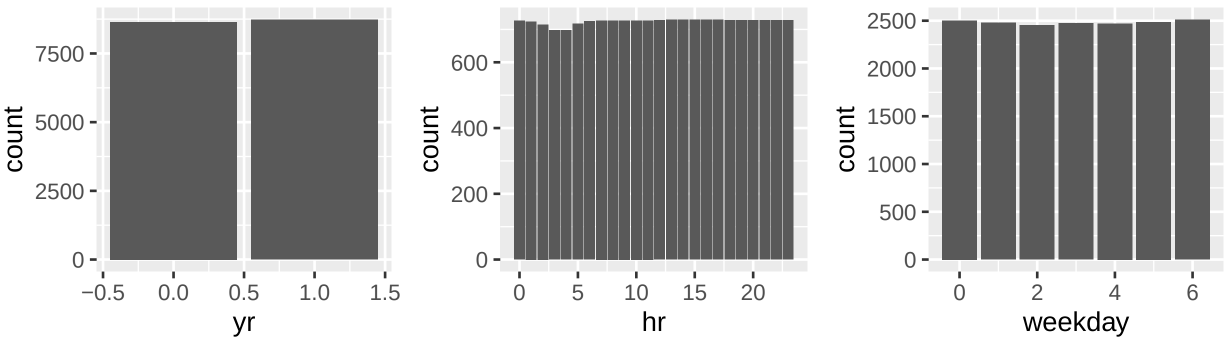


Figure 1: Count of rows per Year, Hour, and Weekday

The distributions of counts for season, month, holiday, and workingday are not interesting on their own, as each category is not meant to have the same number of hours (February has 72 less hours total than March since its got 3 days less). We cannot infer if the dips shown in figure 2 are due to hours containing no rides or just how the categories are set.

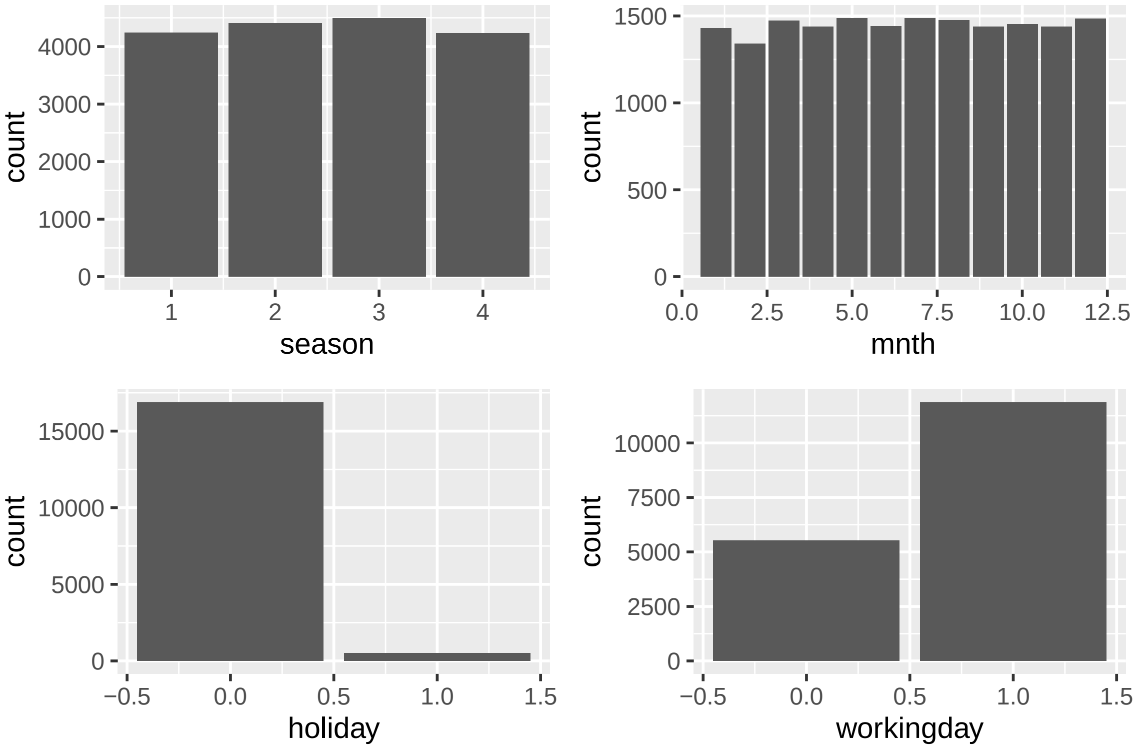


Figure 2: Count of rows per Season, Month, Holiday, and Workingday

Weather related variables

The *weathersit* variable is categorical, where the conditions were classified into four. The data can be taken as ordinal, with '1' being the most 'pleasant' weather, and 4 the least. From the original data source description:

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy;

- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist;
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds;
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.

We see in figure 3 that a majority of cases are classified as ‘1’, and decreasing counts as the weather gets less pleasant. We know the data doesn’t contain an exhaustive list of all hours, but assuming the effect of missing hours is not great, this tells us that the weather is pleasant most of the time in DC. We see that there were only 3 hours where the weather was at its worse and there was at least one bike rental.

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
```

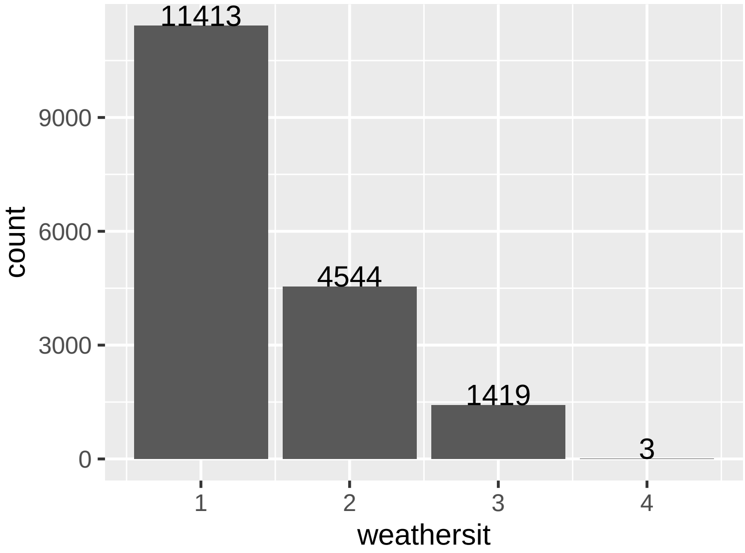


Figure 3: Count of rows per weather category

The rest of the weather related variables, *temp*, *atemp*, *hum* and *windspeed* have been scaled. We’ve transformed the data in new columns to get back original values to help make charts interpretable. In the *temp* plot of figure 4, we see what looks like a symmetric bimodal distribution. From the summary we see that the minimum is -7.06, and the maximum is +39 Celsius, and that the median and mean are similar values, at 15.50 and 15.36, corroborating the general symmetricalness. The feeling temperature,*atemp* shows more of a flattened peak, although there is one value that is recorded about doubly more often than any other. Similarly there’s a few troughs, but generally it seems symmetric, although it has a slightly larger gap between its median and mean. The minimum and maximum here are -16 and +50 Celsius. We would have expected a more continuous distribution, so there might be something going on in regards to rounding or collection of data. Similarly, we would have expected more continuous data for the humidity records, as well as the windspeed.

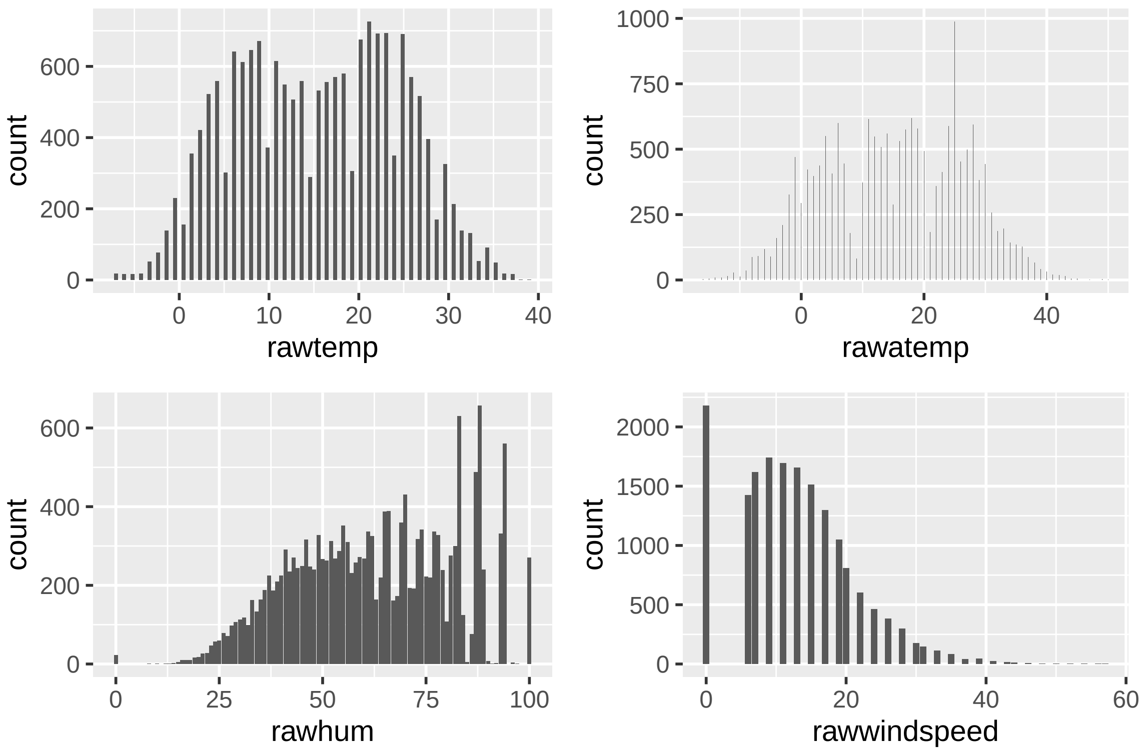


Figure 4: Count of rows per temperature, feeling temperature, humidity, and windspeed

```
summary(df[18:21])
```

```
##     rawtemp         rawatemp          rawhum        rawwindspeed
##  Min.   :-7.06   Min.   :-16.000   Min.   :  0.00   Min.   : 0.000
##  1st Qu.: 7.98   1st Qu.:  5.998   1st Qu.: 48.00   1st Qu.: 7.002
##  Median :15.50   Median : 15.997   Median : 63.00   Median :12.998
##  Mean   :15.36   Mean   : 15.401   Mean   : 62.72   Mean   :12.737
##  3rd Qu.:23.02   3rd Qu.: 24.999   3rd Qu.: 78.00   3rd Qu.:16.998
##  Max.   :39.00   Max.   : 50.000   Max.   :100.00   Max.   :56.997
```

Response data

Finally, *casual*, *registered*, and *cnt* are counts of bikes rented during each 'hour', corresponding to the count of casual users, registered users, and the sum of both. For the count of casual users during a given hour, we see in figure 5 what looks like a steep exponential decline. Intuitively, less users during a given hour happen a lot more often then a lot of users. The distribution of registered users sees less of an extreme drop, with a less steep decline from 50 counts on of individual hours observing a minimum of about 75 rides on.

From the summary, it is really interesting to see that the max amount of casual user rentals in a given hour is 367, whereas that of registered users is 886: more than double! The total amount of rides taken by casual users over the two years sums up to 6.20017×10^5, and that of registered users is 2.672662×10^6. Registered users are accountable for a majority of bike rentals.

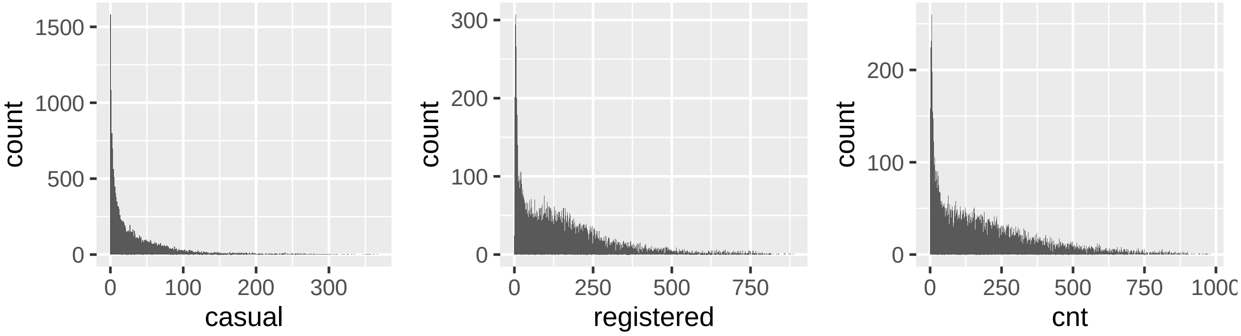


Figure 5: Count of rows per casual user values, registered user values, and total values,

```
summary(df[15:17])
```

```
##      casual         registered         cnt       
##  Min.   :  0.00   Min.   :  0.0   Min.   :  1.0  
##  1st Qu.:  4.00   1st Qu.: 34.0   1st Qu.: 40.0  
##  Median : 17.00   Median :115.0   Median :142.0  
##  Mean   : 35.68   Mean   :153.8   Mean   :189.5  
##  3rd Qu.: 48.00   3rd Qu.:220.0   3rd Qu.:281.0  
##  Max.   :367.00   Max.   :886.0   Max.   :977.0  
```

Correlation among variables

Figure 6 plots a correlation matrix, where we've removed predictors where the factor assigned to it is arbitrary, like *season* and *month*. Variables positively correlated with the total count of bikes rented in an hour are the year, hour of the day, and temperature both measured and felt. Humidity and *weathersit* are both negatively correlated. Variables with near zero correlation are *windspeed*, *holiday*, and *workingday*. It is interesting to see that *workingday* is weakly positively correlated with the count of registered user rentals, whereas more strongly negatively correlated with that of casual users. It is the only variable where the trends are not in the same direction between the two types of users.

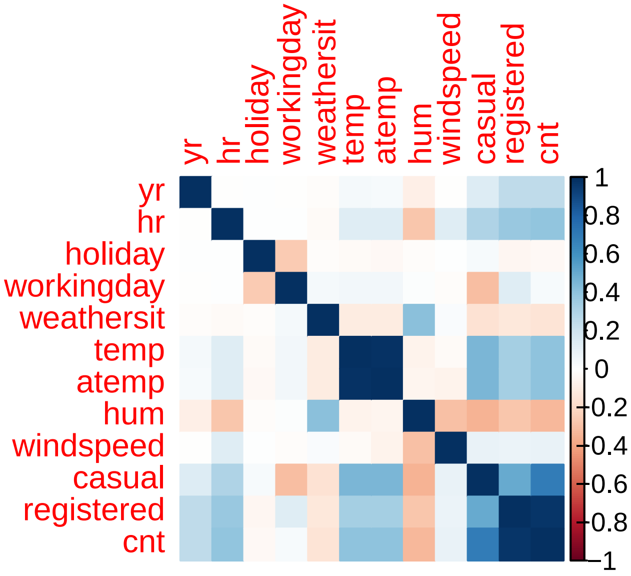


Figure 6: Correlation between select variables

Checking if 'weathersit' and 'holiday' variables are important using ANOVA

The F values greater than 1 and the pvalues less than 0.05 show that both these variables are significant in predicting count of bike rentals, despite their weak correlation.

```
## Analysis of Variance Table
##
## Response: cnt
##                Df    Sum Sq  Mean Sq F value    Pr(>F)
## weathersit      1  11598301 11598301 360.183 < 2.2e-16 ***
## holiday         1    636252   636252  19.759 8.841e-06 ***
## Residuals   17376 559527039    32201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exploratory Charts

Bike Rental Count per Year

The number of bikes rented out has increased from 2011 in 2012, as seen in figure 7. We notice the outlier in 2012 which can be seen as a dot towards the bottom of the plot, and plotted in red in the scatterplot.

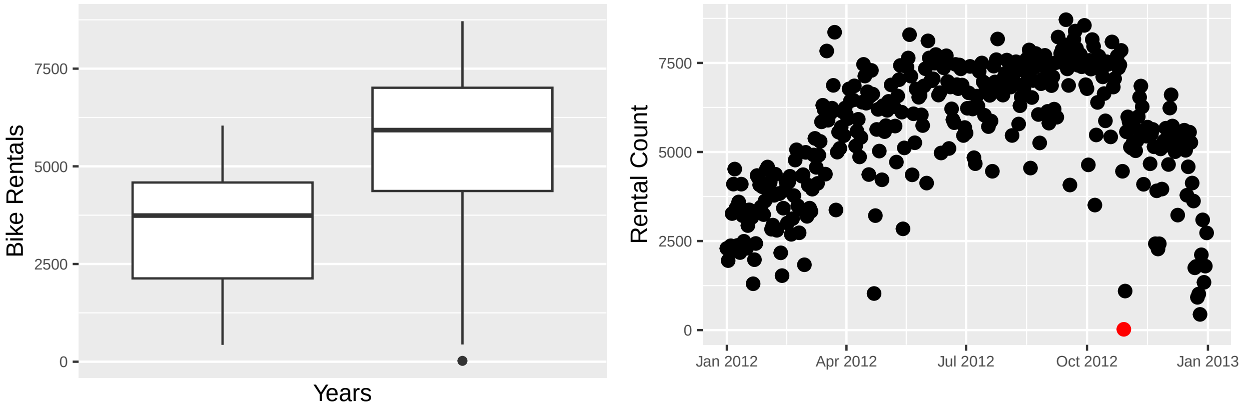


Figure 7: Left: Distribution of bike rental count per year, Right: 2012 daily bike rental count

Let us explore this further:

```
myday_2012 <- subset(day_sum, yr == 2012)
mydata = myday_2012[myday_2012$cnt == min(myday_2012$cnt), c("cnt","dteday")]
cbind(yearly_mean = mean(myday_2012$cnt), mydata)
```

```
##     yearly_mean cnt     dteday
## 668    5599.934  22 2012-10-29
```

We see that the number of bikes rented out on October 29, 2012 was underwhelmingly lower than the yearly average for 2012 making it an outlier. We investigated further to see why that is and found that that is the day Hurricane Sandy landed on the east coast.

Bike Rental Counts per Month

Figure 8 plots the count of bikes used for every month, split by year. Interestingly, 2011 sees a flat peak in June and a gradual decline into the winter months. The next year has a big jump from February to March, then climbs gradually to peak in September before seeing a steeper decline into the winter months.

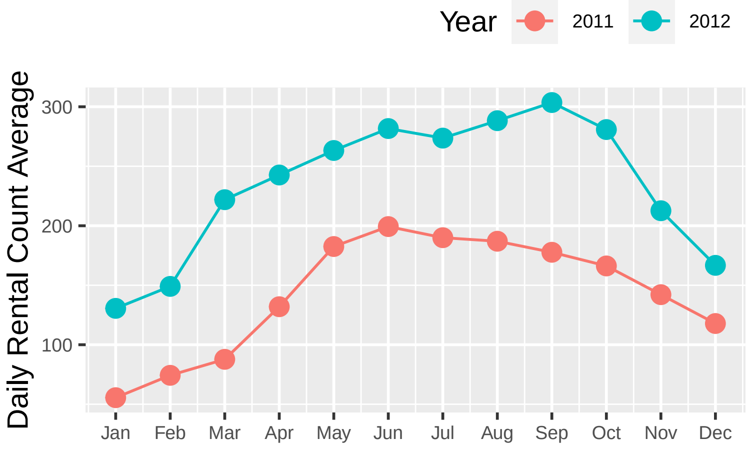


Figure 8: Average daily bike rentals per month by year

Bike Rental Counts per Hour

Next, in figure 9, we can see that both years follow a similar trend in average rental count per hour of day. There is a similar low number of bikes rented out in the hours of 2,3,4 and 5 in both years. The change is greatest in the daytime hours: 7am to 9pm. In an seasons, we observe peaks at 8am and 5pm, suggesting that people use the rental system as transportation to work.

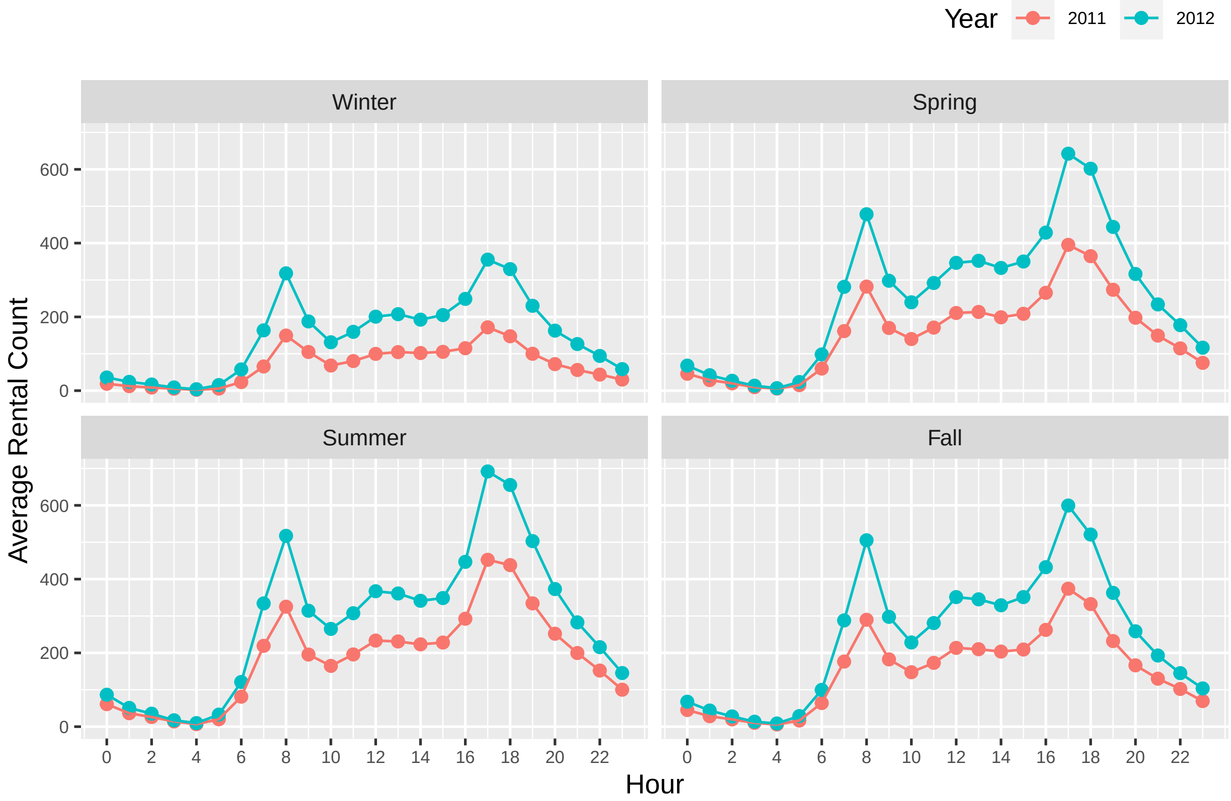


Figure 9: Average bike rental counts per hour, per season and by year

Bike Rental Counts per Weekday

In figure 10, we interestingly don't observe a drop of rides during the weekends in every season. The trends are not similar for neither both years nor all seasons.

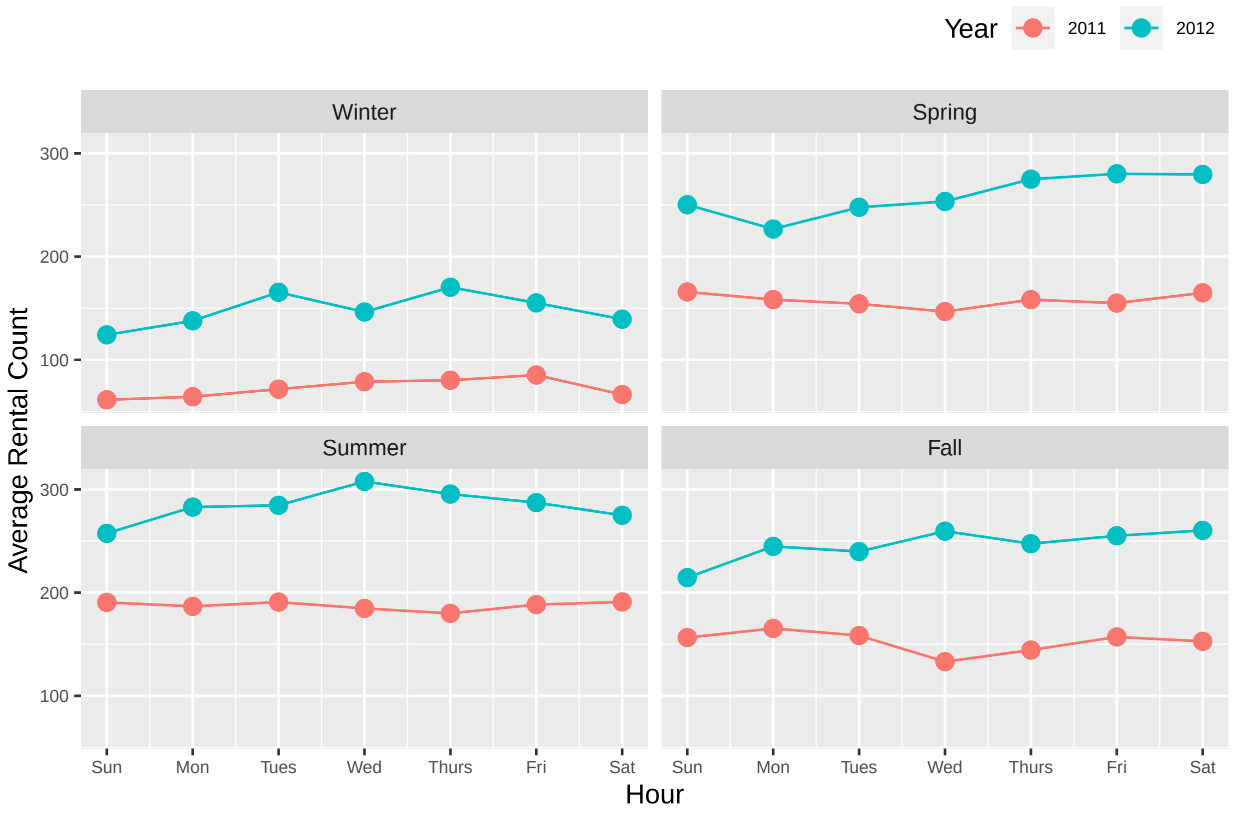
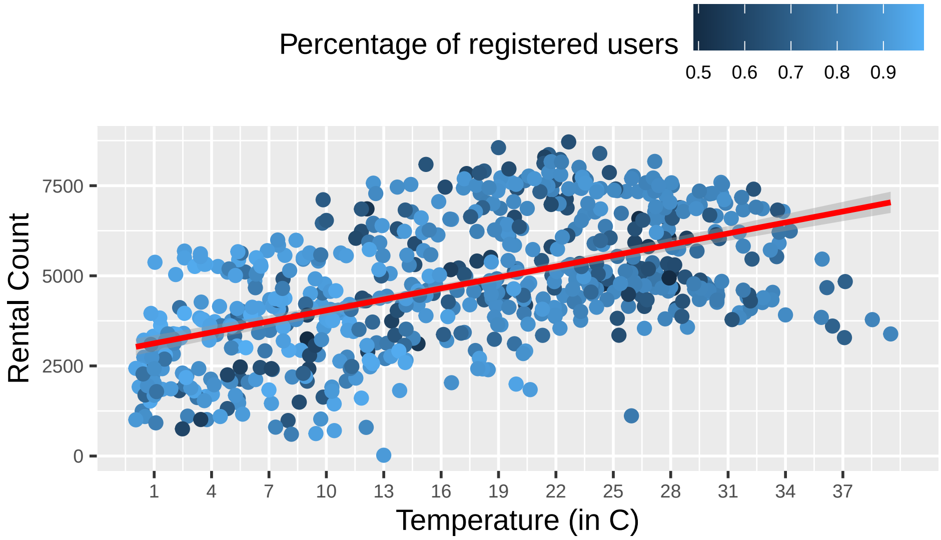


Figure 10: Average bike rental counts per weekday, per season and by year

Bike rental count and temperature

In figure 11, We see that generally, the number of bike rentals per day increases with the temperature. From the colour scheme, we see that the points for the lower temperatures are generally lighter, showing that a greater proportion of the trips taken in that day are by registered users. There is never a day where casual users make up more than 50% of the trips taken, regardless of temperature.

\begin{figure}[H]



{

}

\caption{Daily count of bike rentals vs daily temperature, colored by % of registered users} \end{figure}