**Report BiKCitY**

In this report, we will analyze the data collected in the file “DatasetBikeSharing.xlsx”, an Excel file containing information about bike sharing. Our objective will be to use the information that we can extract from the data (using statistical methodologies) in order to propose a strategy to the company BiKCitY on how to increase the number of rented bikes.

First, we need to give a general look at the dataset. The variables in the dataset are the following:

1. season: Winter, Spring, Summer, Fall;
2. holiday: not a holiday, a holiday;
3. workingday: not working, working;
4. temp: temperature;
5. atemp: feeling temperature;
6. hum: humidity;
7. windspeed: wind speed;
8. cnt: count of total rental bikes.

Before we can start with the analysis, we will need to inspect the data looking for outliers, missing data and inconsistencies.

We will start by looking for missing data. Using our statistical software, we can confirm that there are not missing values in any of the variables in the dataset (both in the numerical and in the categorical variables).

Next, we will look for outliers in the numerical variables. We can do it using the box and whiskers plot.

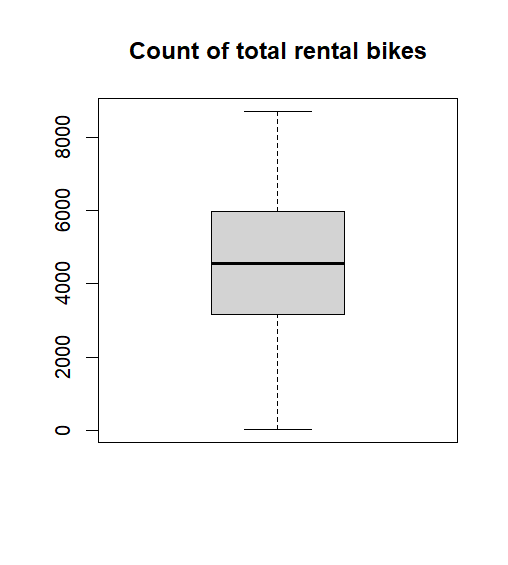
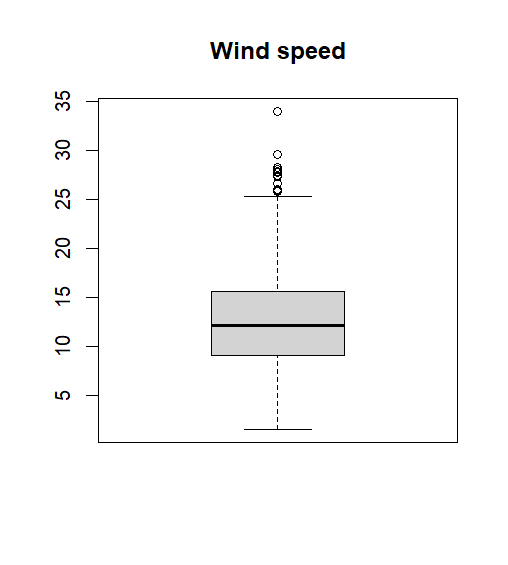
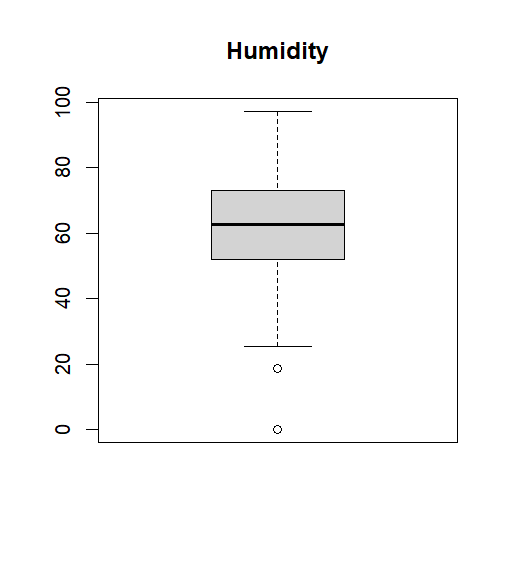
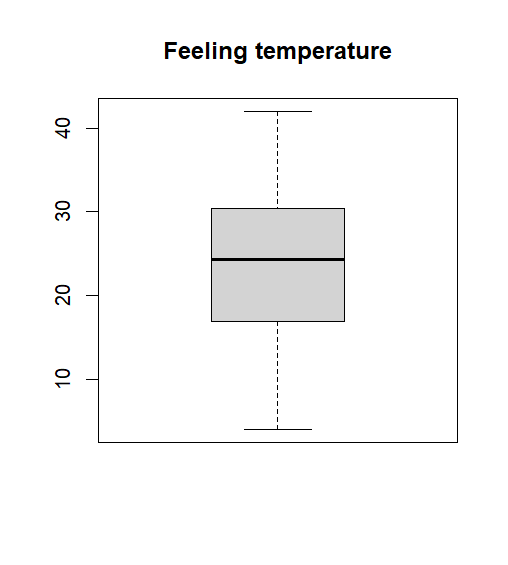
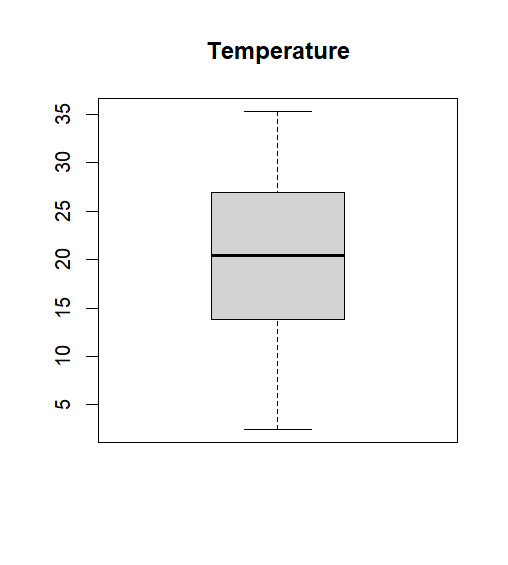


Figure 1 - Boxplots of the variables

Looking at the boxplots in Figure 1, we can conclude that the variables “Temperature”, “Feeling temperature” and “Count of total rental bikes” do not contain outliers. However, the variables “Humidity” and “Wind speed” seem to contain some. In order to confirm if those values can be considered outliers, we can look at the histograms of the two variables.

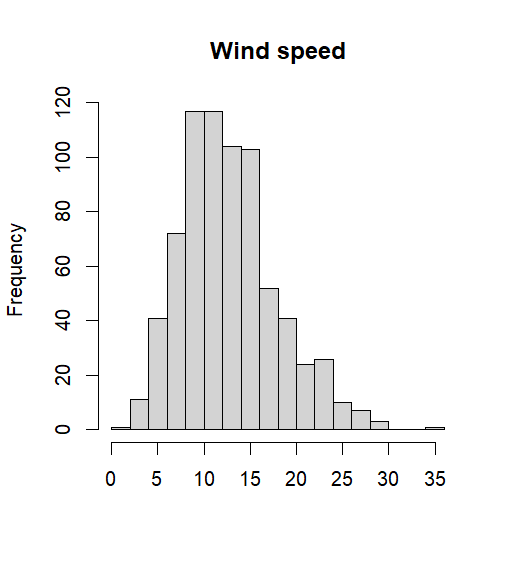


Figure 2 - Histogram of Wind speed

First, we will look at the histogram of “Wind speed” (Figure 2). It can be easily observed that the values that in the box and whiskers plot could be thought as outliers are, actually, due to the positive asymmetry of the distribution. So, these values are coherent with the distribution and cannot be considered as outliers.

Now, we can look at the histogram of Humidity (Figure 3). It can be observed that the value immediately below the whisker in the boxplot cannot be considered an outlier. In fact, it is coherent with the distribution of the values of humidity and is due to a higher variability then expected (the whiskers are either the min and max value or built in an arbitrary way).

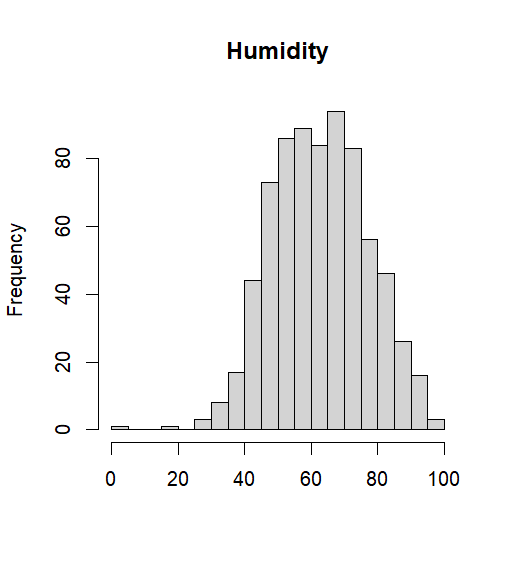


Figure 3 - Histogram of Humidity

However, it can also be observed that the value on the extreme left (which is located in the 69th row) of the distribution is indeed an outlier. In fact, it is not coherent with the distribution of the values of the variable. So, we will need to do something about this value.

We can treat the outlier by deleting it.

Lastly, we have to check for inconsistencies in the data. Looking at the graphs in Figure 1, it can be easily observed that there are not any inconsistencies in those variables. However, we also have to check for inconsistencies in the categorical variables. We can do it by looking at the pie charts of the variables.

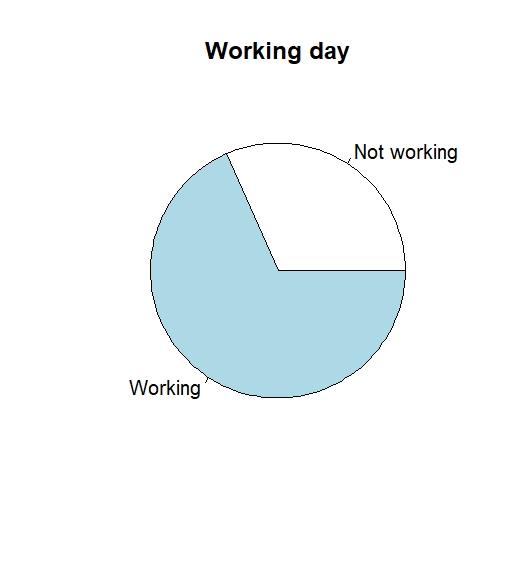
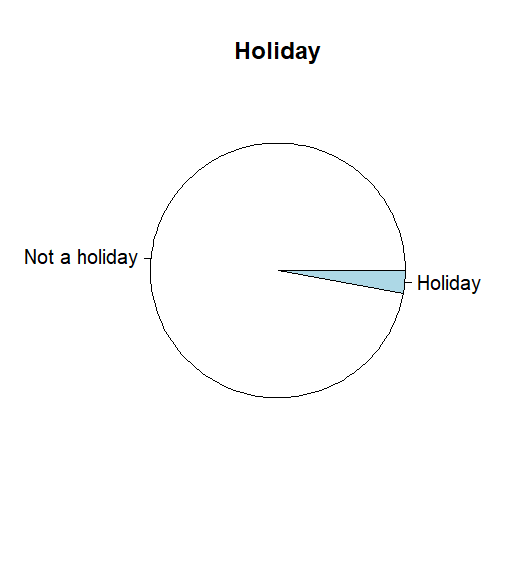
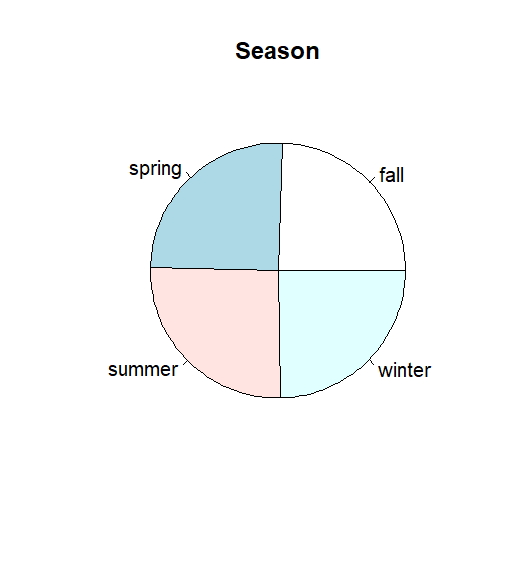


Figure 4 - Pie charts of the categorical variables

It can be easily ascertained from the graphs that there are not any inconsistencies in the categorical variables as all of the values fall in one of the previously described categories.

Now, we can proceed with the analysis of the dataset. In order to identify which dimensions are the most impactful on the number of rented bikes, we will compute linear regression using the variables in the dataset. Before doing that, it can be useful to look at the different plots with “cnt” on the y axis in order to have a better understanding about the relationship between each of the other variables (which will be used as predictors) and the count of total rental bikes.

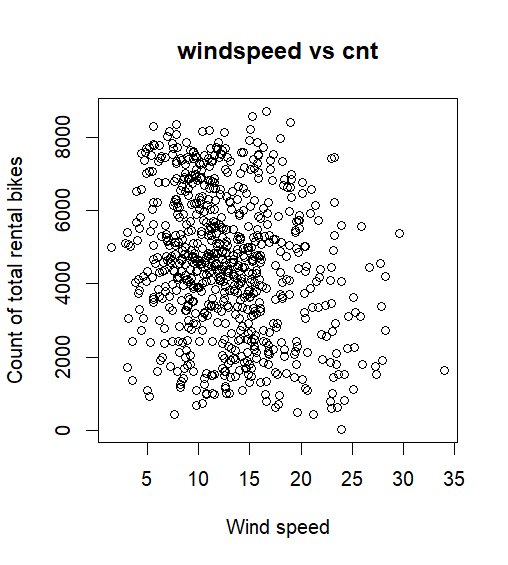
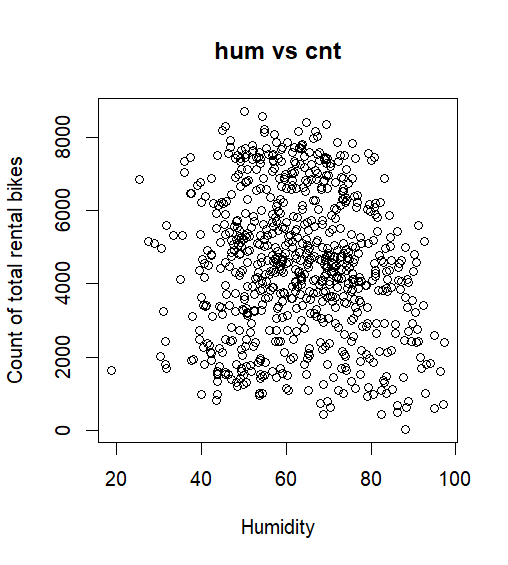
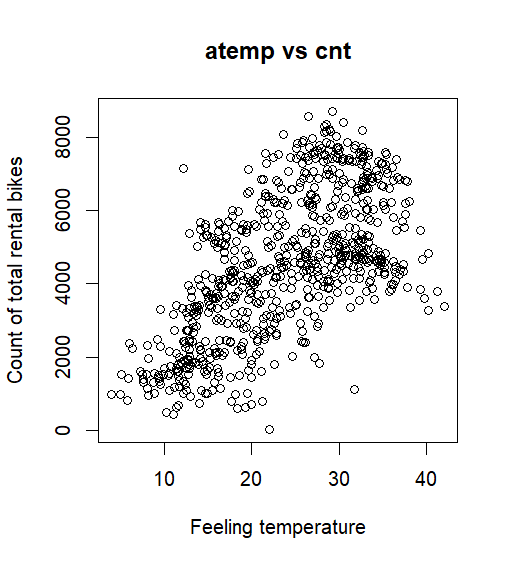
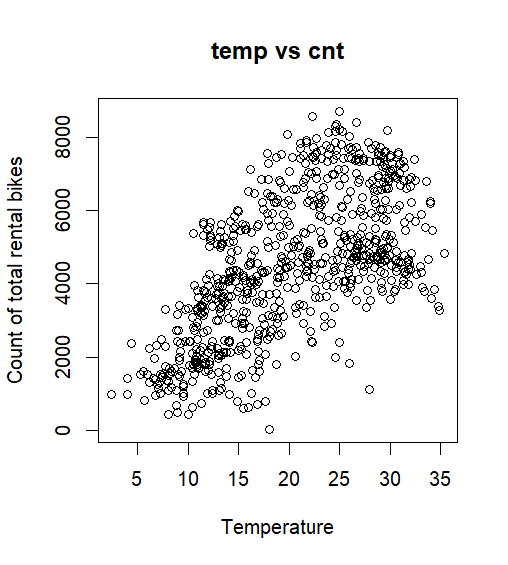
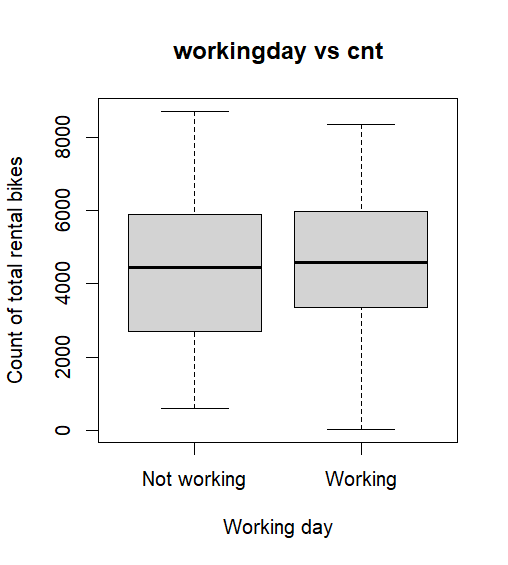
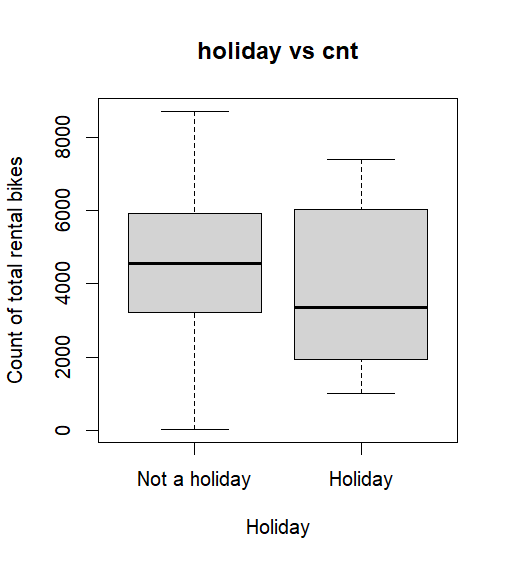
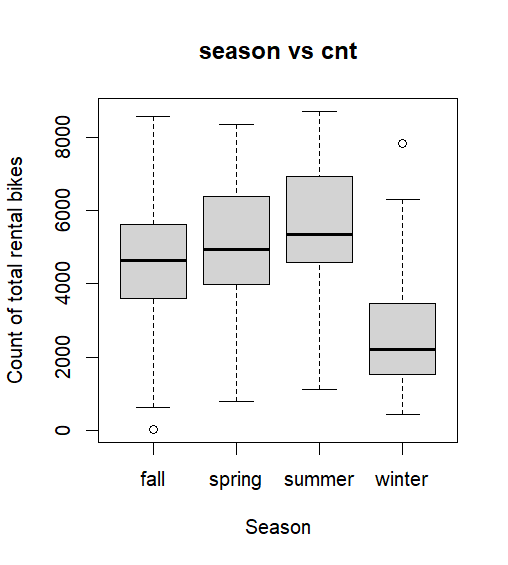


Figure 5 - Graphs with cnt on y axis

Looking at the graphs in Figure 5, it can be observed that the variables “season”, “temp” and “atemp” seem to be associated with “cnt” (the values of cnt seem to change when one of the previously described variables changes). Also, the points in the scatterplots of the numerical variables seem to be aligned around a straight line. So, we can conclude that there will not be the need to consider polynomial regression models in the following analysis. The other four variables (“holiday”, “workingday”, “hum”, “windspeed”) do not seem to be associated with the count of total rental bikes or it is too difficult to identify the specific association only looking at the graphs. So, we will include these variables in the models to ascertain their effect on the response “cnt”.

We can now proceed with the computation of the linear regression model. We will start with a model that includes all of the variables then we will apply the stepwise selection approach in order to identify the best possible subset of predictors.

Using our statistical software, we can easily compute what we have just described. The estimated coefficients for each variable (which describe the mean change in the response following a unit variation of the corresponding predictor for numerical variables and the mean difference of the response between the categories and the comparison category for categorical variables) are the following (S = significative; NS = the coefficient cannot be considered different from zero):

1. intercept: 5218.892 (S);
2. season - spring: -576.640 (S);
3. season - summer: -1090.068 (S);
4. season - winter: -1426.057 (S);
5. holiday - Holiday: -554.645 (NS);
6. workingday - Working: 80.403 (NS);
7. temp: 130.138 (S);
8. atemp: 28.799 (NS);
9. hum: -39.765 (S);
10. windspeed: -62.684 (S).

It can be observed that some of the coefficients cannot be considered different from zero. This fact implies that the impact of these predictors is not significative on the response which means:

* for numerical variables: the variation of the predictor does not affect the response;
* for categorical variables: the mean value of the response is approximately the same for the considered category and the one used as a comparison.

Now, we can proceed with the identification of the best possible model using the stepwise selection approach. After the computation of the procedure, the new coefficients are the following:

1. Intercept: 5326.672 (S);
2. season - spring: -581.089 (S);
3. season - summer: -1113.743 (S);
4. season - winter: -1429.778 (S);
5. holiday - Holiday: --617.093 (S);
6. temp: 162.051 (S);
7. hum: -39.641 (S);
8. windspeed: -63.924 (S).

We can observe that the variables “Working day” and “Feeling temperature” were removed from the model. Also, the removal of these variables caused the reduction of the p-value of the variable “holiday - Holiday” which is now significative. We can conclude that these are the best predictor for the number of rented bikes.

**Conclusion**: we will now look at the results of the last model to propose a strategy to BiKCitY. Looking at the coefficients from the model, we can say that the most count of total rental bikes is, on average, higher in Fall than in any other season, it is lower during holidays, it increases when the temperature rises and it lowers when the humidity and the windspeed rise.

We can separate between calendar variables and climatic variables:

Calendar variables: it can be observed that the most bikes are rented in the Fall and in days that are not holidays. In order to increase the rentals during the other time periods, one possible strategy would be to decrease the price of the service during holidays and the other three seasons.

Climatic variables: it can be observed that the rentals decrease when the humidity and the windspeed rise. So, one possible strategy would be to offer free clothes to contrast humidity and wind (for example, thermal coats) in days with high levels of humidity and wind speed.