Washington Bike Rental Data

Introduction

The data show bike rentals in Washington in the years 2011 and 2012. The bike rentals are differentiated between registered and casual users.

These data are combined with weather data and common information like holidays and weekday. The weather data provided are a category of weather situation like sunny or heavy rain, temperature, felt temperature, humidity and windspeed.

Exploratory Data Analysis

The project has a total of 17 379 data points. There are no missing values.
Some insights are shown in the diagrams on the following pages. The most important aspects are:

- Bike rentals drop to 50 % if it is raining or snowing compared to clear weather (see Figure 1).
- In winter month people are less sensitive to slightly bad weather, as rentals are equal to those at clear weather (see Figure 2).
- People are also more tolerant to bad weather in summer or at night (see Figure 2+3).
- Bike rentals rise until very hot temperatures, then they drop slightly (see Figure 4).
- Bike rentals are highest between 0.1 and 0.4 humidity and then drop steadily (see Figure 5).
- In winter only about 10 % of users are casual users whereas up to 40 % of users are casual users in summer (see Figure 6).
- Most bikes are rented in the morning (at 8 am) and evening (5 to 6 pm) independent of month, people probably use the bikes most to get to or from work (see Figure 7).
- Nearly no bikes are rented in the night independend of month (see Figure 7).
- Bike rentals of casual users are opposit to those of registered users, the highest percentage of casual users is in the night and midday (see Figure 8).

Forecast Model

To forecast the hourly bike rentals, a Random Forest model was chosen. A Random Forest is a collection of lots of decision trees. Each decision tree is different from the others. Combining the trees gives better results as weaknesses of each tree are compensated by many others.

With this model, further information than just the time series can be taken into account. This is needed, as the weather data is highly important for this task.

The data is split into train and test set. As it is a time series, the last 25 % of data points are used as a test set.

The mean absolute deviation is 18 bike rentals. At mean total bike rentals of 189 this is an error of 10 %. Smaller values of bike rental counts tend to be overestimated by the model, meaning that the residue is negative (see Figure 9).

Diagrams

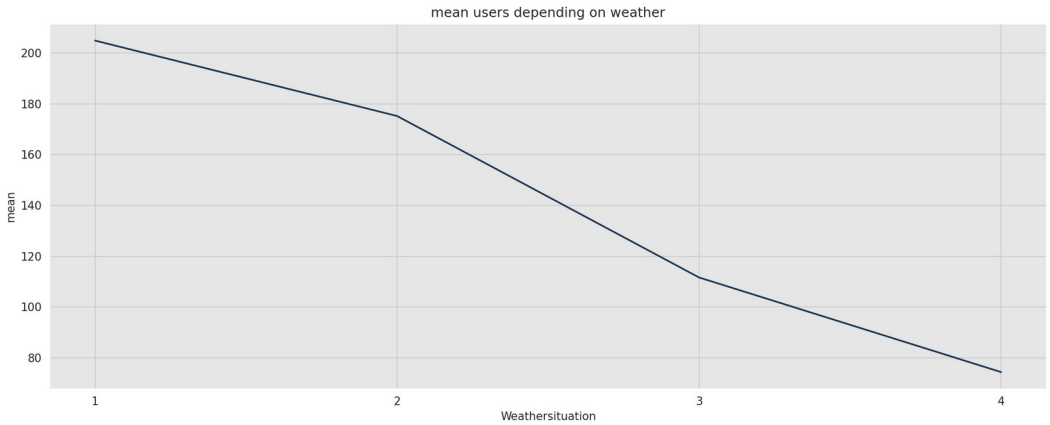

Figure 1: Mean count of users depending on the weather situation from 1: clear weather to 4: heavy rain

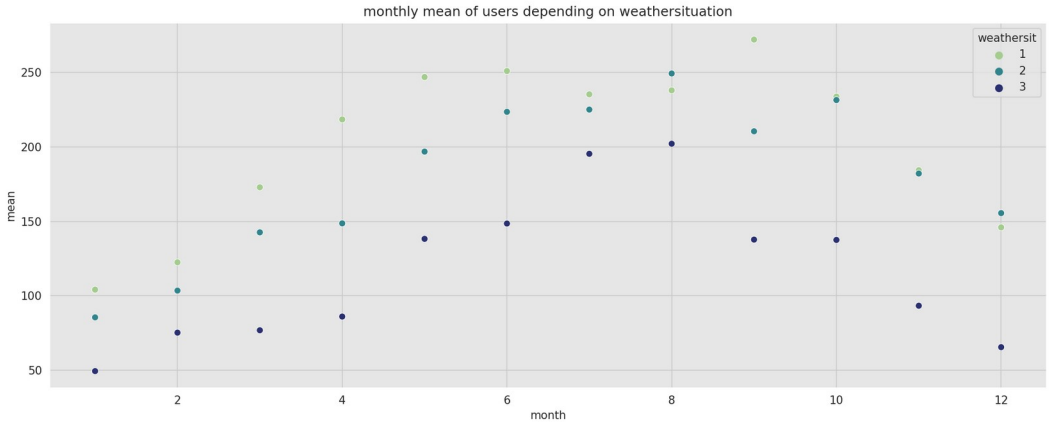

Figure 2: Monthly users depending on weather situation. Weather situation 4 exists only a very few times, therefore it is not shown here.

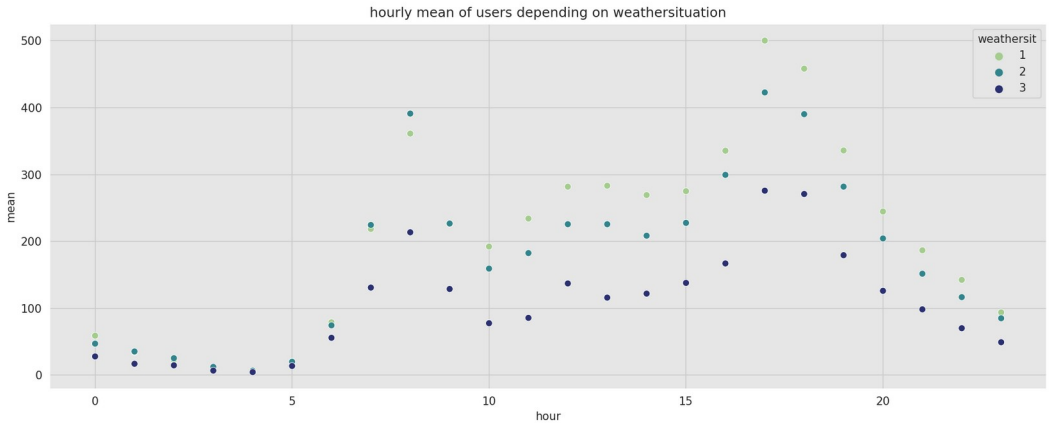

Figure 3: Hourly mean of users depending on weather situation. Weather situation 4 exists only a very few times, therefore it is not shown here.

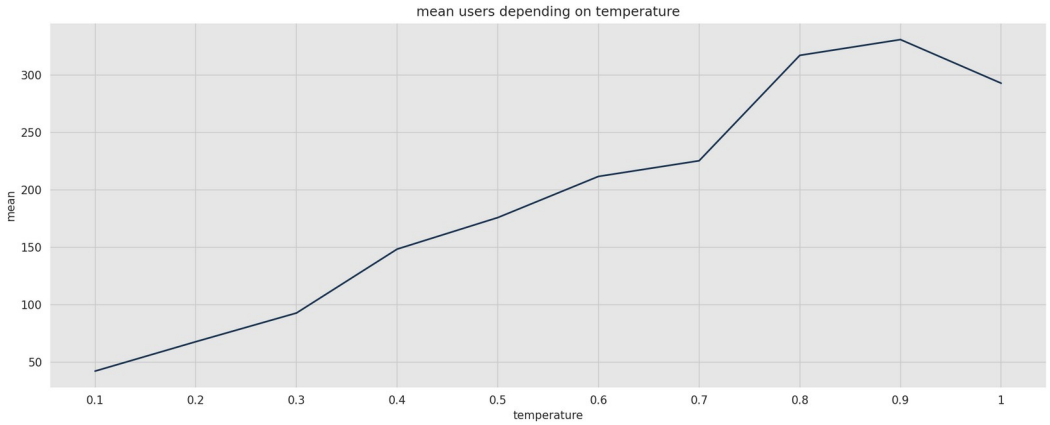

Figure 4: Mean users depending on temperature. The temperature is normalized to the maximum temperature.

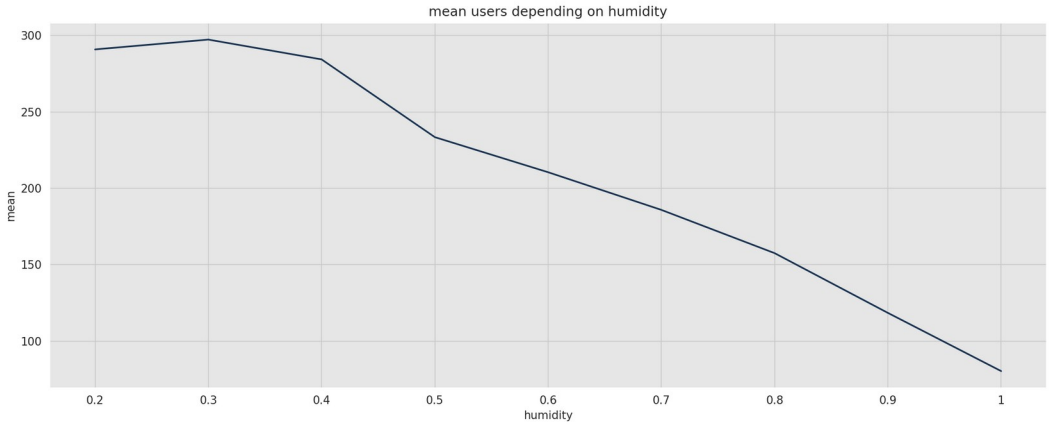

Figure 5: Mean users depending on humidity

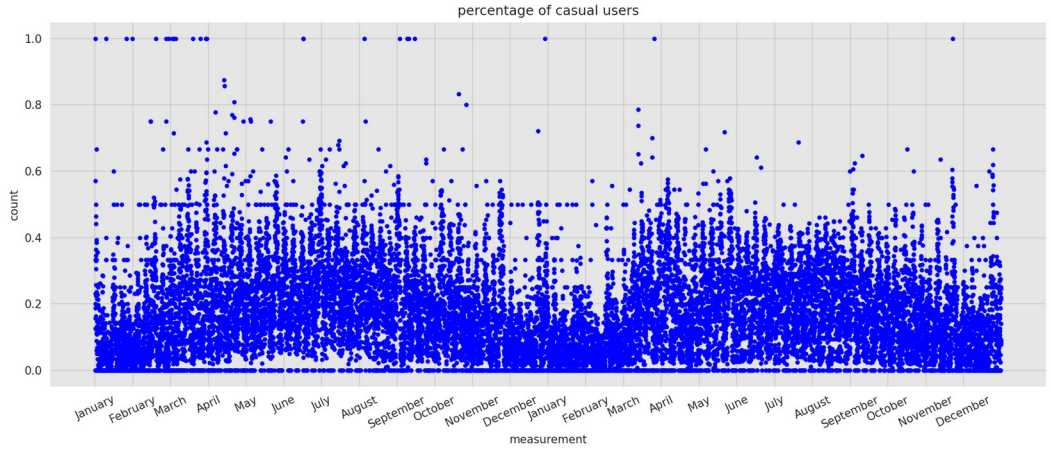

Figure 6: Percentage of casual users compared to registered users for the two year period 2011 to 2012

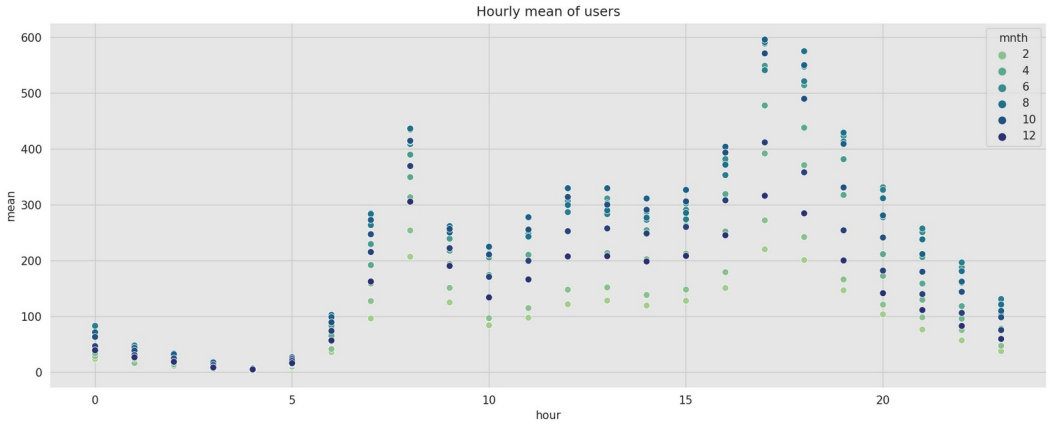

Figure 7: Hourly mean of users for each month

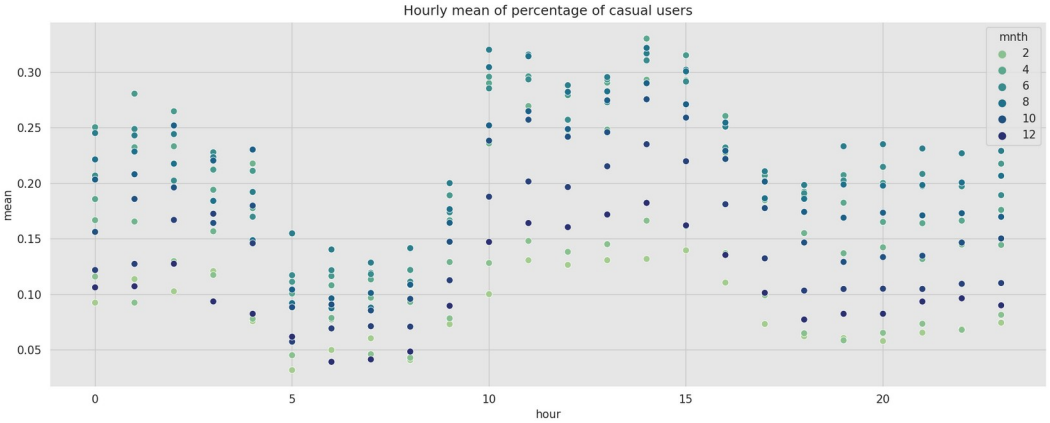

Figure 8: Hourly mean of the percentage of casual users per month

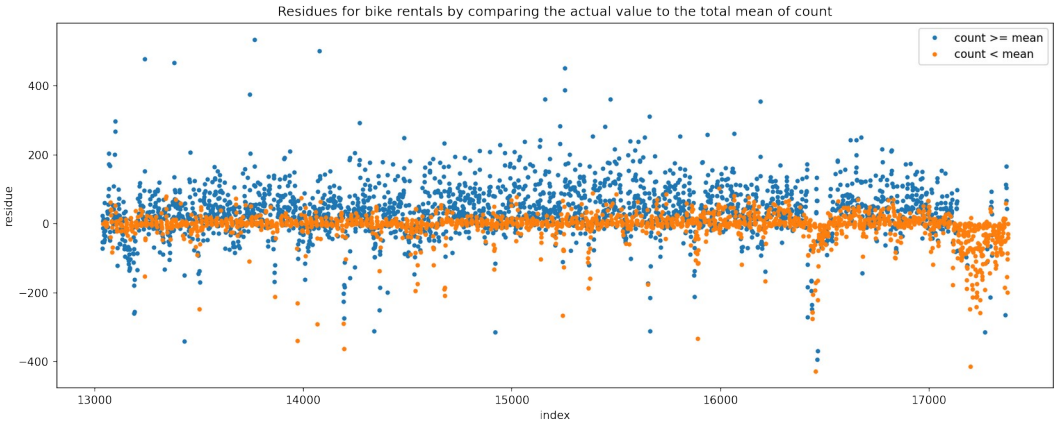

Figure 9: Residue of the predictions by a Random Forest model