Analysis of a spatio-temporal dataset

Algorithmic Data Analysis – Coding Assignment 5

Giulia Ortolani

March 27, 2024

Resources: See *References* section
Collaborations: No collaborations

1 Dataset presentation

The Bike Sharing dataset[1] contains the hourly and daily count of rental bikes between years 2011 and 2012 in the Capital Bikeshare system (Washington D.C., USA), alongside corresponding weather and seasonal information [1]. This dataset offers valuable insights into bike-sharing systems, which represent modern alternatives to traditional bike rentals. These systems enable users to conveniently rent a bike from one location and return it to another, contributing significantly to urban transportation, mitigating traffic congestion, addressing environmental concerns, and promoting healthier lifestyles.

The dataset is particularly focused on understanding how various factors such as weather, season, and time of day impact bike rentals. It encompasses two years of historical data (2011 and 2012) from the Capital Bikeshare system in Washington D.C., USA, collected at both two-hour intervals and daily intervals. Additionally, the dataset includes supplementary information about weather conditions obtained from reliable weather websites.

The dataset contains 731 observations of 15 variables. These variables are presented and explained in Table 1.

1.1 Explorative analysis

From the initial explorative analysis of the dataset, we found no duplicated values or missing data (NA), ensuring the dataset's integrity. The maximum number of rented bikes was 8714 on September 15, 2012, while the minimum was only 22 on October 29, 2012, coinciding with Hurricane Sandy.

In terms of temperature, the actual air temperature ranged from 2 to 35 degrees Celsius, and the perceived air temperature ranged from 4 to 42 degrees Celsius. Humidity is indicated as percentage and varies from 0 to 97, with 0 likely representing a measurement error rather than absolute zero.

Wind speed data showed a maximum of 34 meters per second and a minimum of 1.5 meters per second.

1.1.1 Data visualization

I created plots to visualize the number of rented bicycles over time, both at each timestamp (Figure 1) and their moving average with a 30-day intervals (Figure 2). Notably, there was an increase in bike rentals in 2012 compared to 2011, indicating potential growth in bike-sharing usage over the years.

[1] `https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset`

Variable	Description
instant	Record index
dteday	Date
season	Season (1:spring, 2:summer, 3:fall, 4:winter)
yr	Year (0:2011, 1:2012)
mnth	Month (1 to 12)
hr	Hour (0 to 23)
holiday	Weather day is holiday or not
weekday	Day of the week
workingday	1 if day is neither weekend nor holiday, otherwise 0
weathersit	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
temp	Normalized temperature in Celsius (divided by 41)
atemp	Normalized feeling temperature in Celsius (divided by 50)
hum	Normalized humidity (divided by 100)
windspeed	Normalized wind speed (divided by 67)
casual	Count of casual users: individuals who use the bike rental service occasionally or on a one-time basis
registered	Count of registered users: individuals who have signed up for a membership or subscription with the bike rental service
cnt	Count of total rental bikes (casual + registered)

Table 1: Description of Variables of the Dataset

1.1.2 Correlation plot

I also generated a correlation plot (Figure 3) to examine the relationships between different variables. I observed high correlation values between `weathersit` and `hum`, as well as with `cnt`, `temp`, and `atemp`. This indicates that weather conditions significantly influence bike rental counts and temperature-related variables. More in details, these correlations may suggest:

- Correlation between `weathersit` and `hum`: high humidity levels might be associated with certain weather situations (e.g., rain, fog) captured by the weathersit variable. This correlation suggests that certain weather conditions tend to coincide with higher humidity levels.

- Correlation between `weathersit` and `cnt`: the total rental bike count (`cnt`) is strongly correlated with weather conditions (`weathersit`). This indicates that weather plays a crucial role in determining the demand for bike rentals. For example, adverse weather conditions such as rain or snow might lead to a decrease in bike rentals, while clear weather might result in higher demand.

- Correlation between `weathersit` and `temp`: temperature (`temp`) is closely related to weather conditions (`weathersit`). This is intuitive as weather situations like sunny or cloudy weather are often associated with specific temperature ranges.

- Correlation between `weathersit` and `atemp`: similarly, the feeling temperature (atemp) is correlated with weather conditions (weathersit). This suggests that people's perception of temperature is influenced by the prevailing weather conditions.

Conversely, we found a low correlation between `workingday` and `casual`, suggesting that casual rentals are more likely on weekends, while individuals commuting to work by bike are more likely to be registered users.

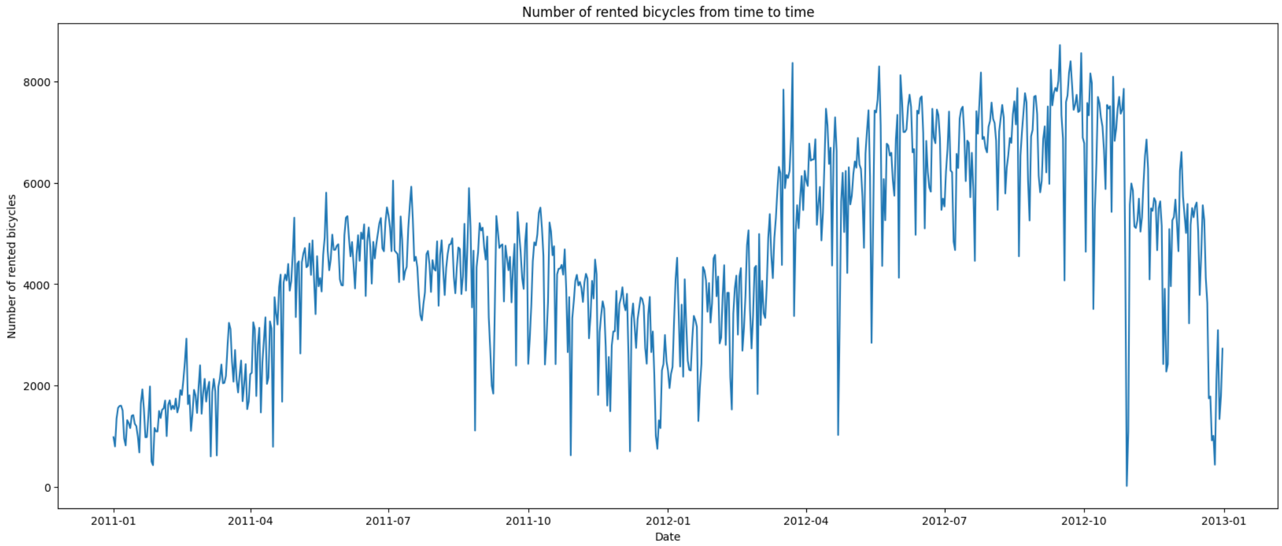

Figure 1: Number of rented bicycles at each timestamp

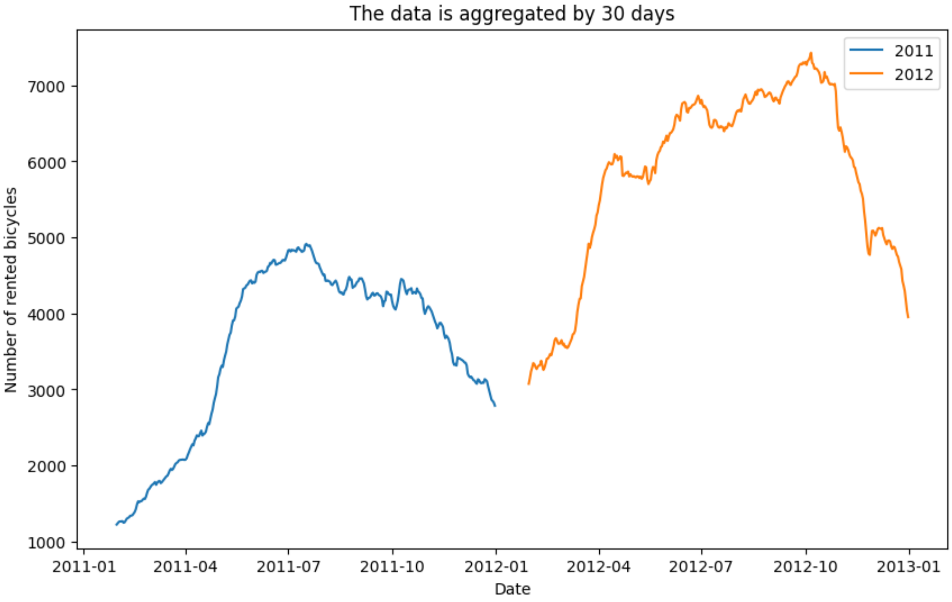

Figure 2: Number of rented bicycles (moving average with 30-day intervals)

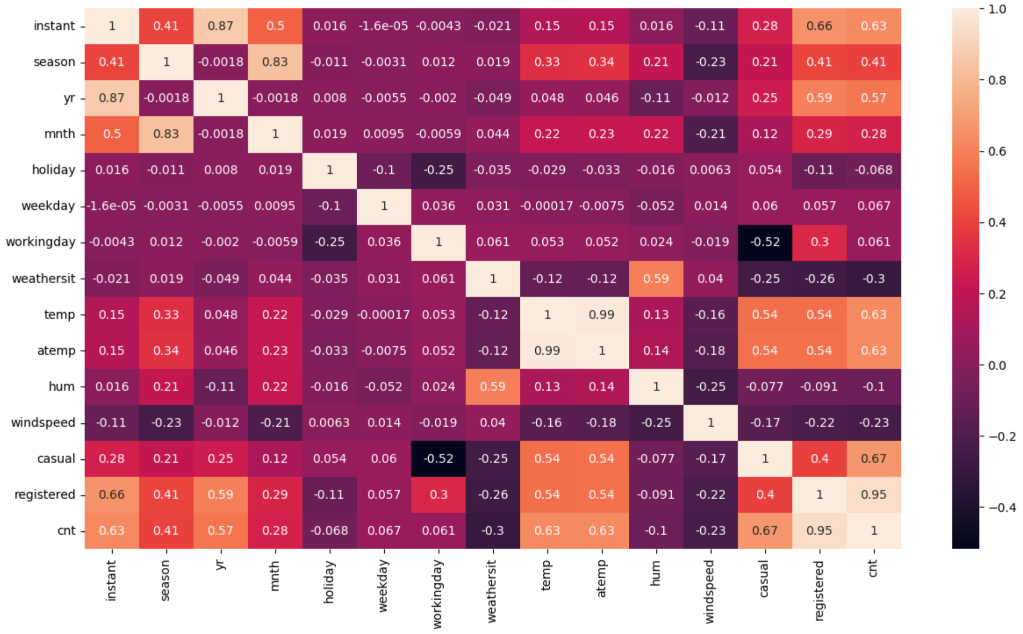

Figure 3: Correlation plot

2 Isolation Forest

Let's run an anomaly detection algorithm on this dataset to identify anomalies in bike rental counts based on weather conditions. Since we have observed high correlations between weather conditions and rental counts, anomalies in rental counts during certain weather situations could indicate unusual or unexpected patterns that deviate from the norm.

Anomaly detection helps to identify unusual patterns in bike rental counts that may occur during specific weather conditions, such as unexpectedly low or high rental counts during certain weather situations. By detecting these anomalies, bike sharing operators or system managers can investigate the underlying causes, such as equipment malfunctions, service disruptions, or unusual weather events, and take appropriate actions to address them, such as adjusting service operations, improving maintenance procedures, or implementing contingency plans.

2.1 Implementation

We select features relevant to the anomaly detection process, including `cnt` (total rental bike count), `weathersit` (weather situation), `hum` (humidity), `temp` (temperature), and `atemp` (feeling temperature).

In the code, first we need to create a subset of the dataset containing only the selected features. Then, we initialize an Isolation Forest model with a specified contamination parameter (indicating the expected proportion of anomalies in the dataset) and fit it to the subset of data.

We use the trained Isolation Forest model to predict anomalies in the dataset and we add an anomaly column to the dataset to indicate whether each data point is an anomaly (-1) or not (1).

We tried different values for the contamination level: in Figure 4 we have four graphs, each representing anomaly detection results using different levels of contamination in the Isolation Forest algorithm. Varying the contamination level allows us to explore how different proportions of outliers affect the detection of anomalies in bike rental counts. This analysis helps us understand the sensitivity of the anomaly detection process to outlier proportions.

From this analysis we can observe that the outliers consistently fall ”below” the average. This suggests that they consistently represent anomalies where the number of rented bikes is lower than the average during those periods.

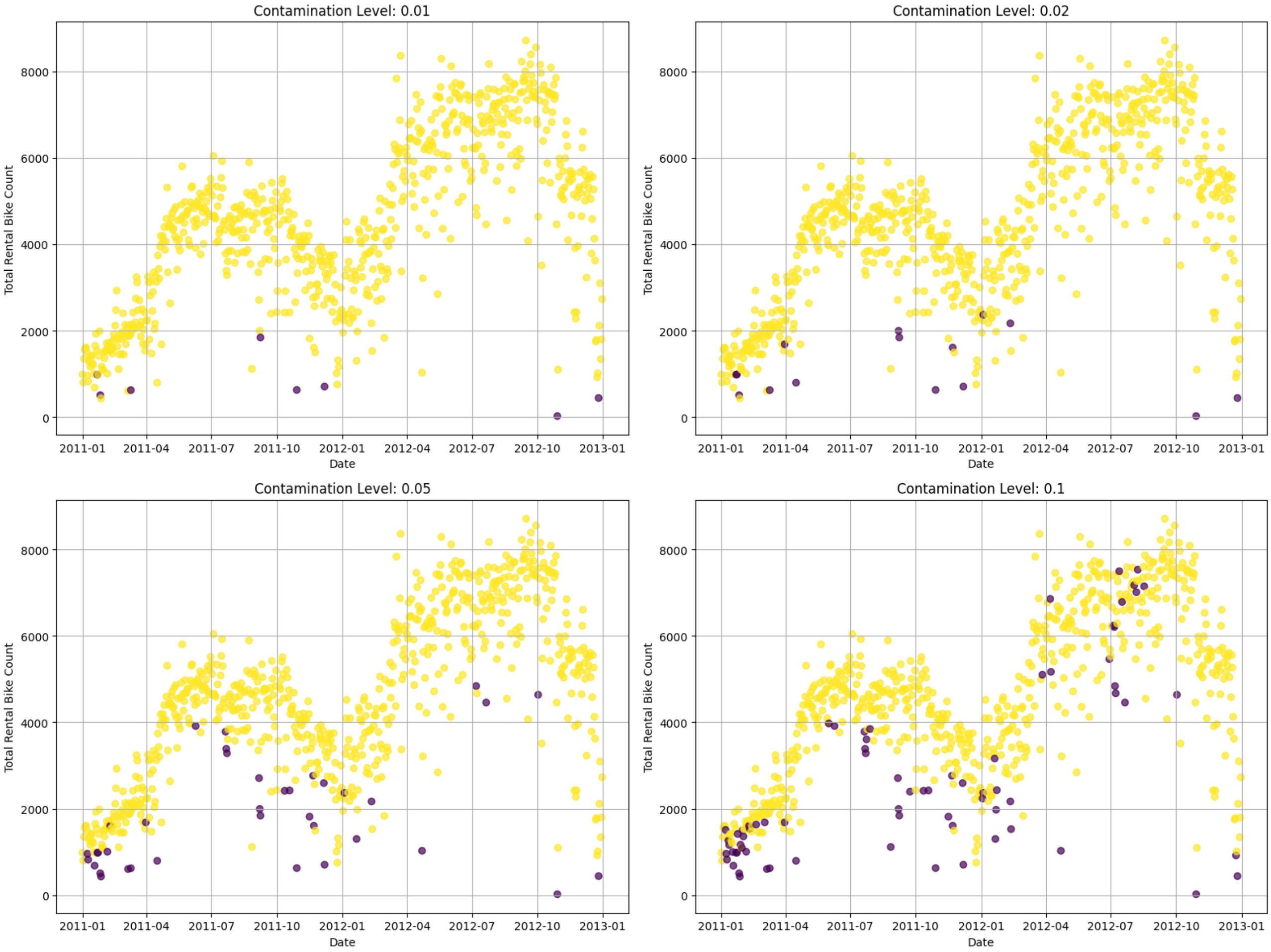

Figure 4: Anomaly detection

2.1.1 Year-wise Anomaly Analysis

To conduct a more precise analysis, let's run the anomaly detection algorithm separately for each of the two different years in the dataset. By doing so, we aim to gain deeper insights into the patterns and anomalies specific to each year. This approach accounts for potential variations in bike rental behavior, weather conditions, and other factors that may differ between the two years.
We expect this separate analysis to reveal year-specific anomalies and patterns that may not be apparent when analyzing the entire dataset together. For instance, it may uncover seasonal trends, changes in rental demand over time, or the impact of specific events or weather conditions unique to each year. By isolating the data for each year, we can better understand the dynamics of bike rental activity and identify anomalies more accurately within the context of each individual year. The contamination parameter chosen in this case was 0.05.
In Figure 5

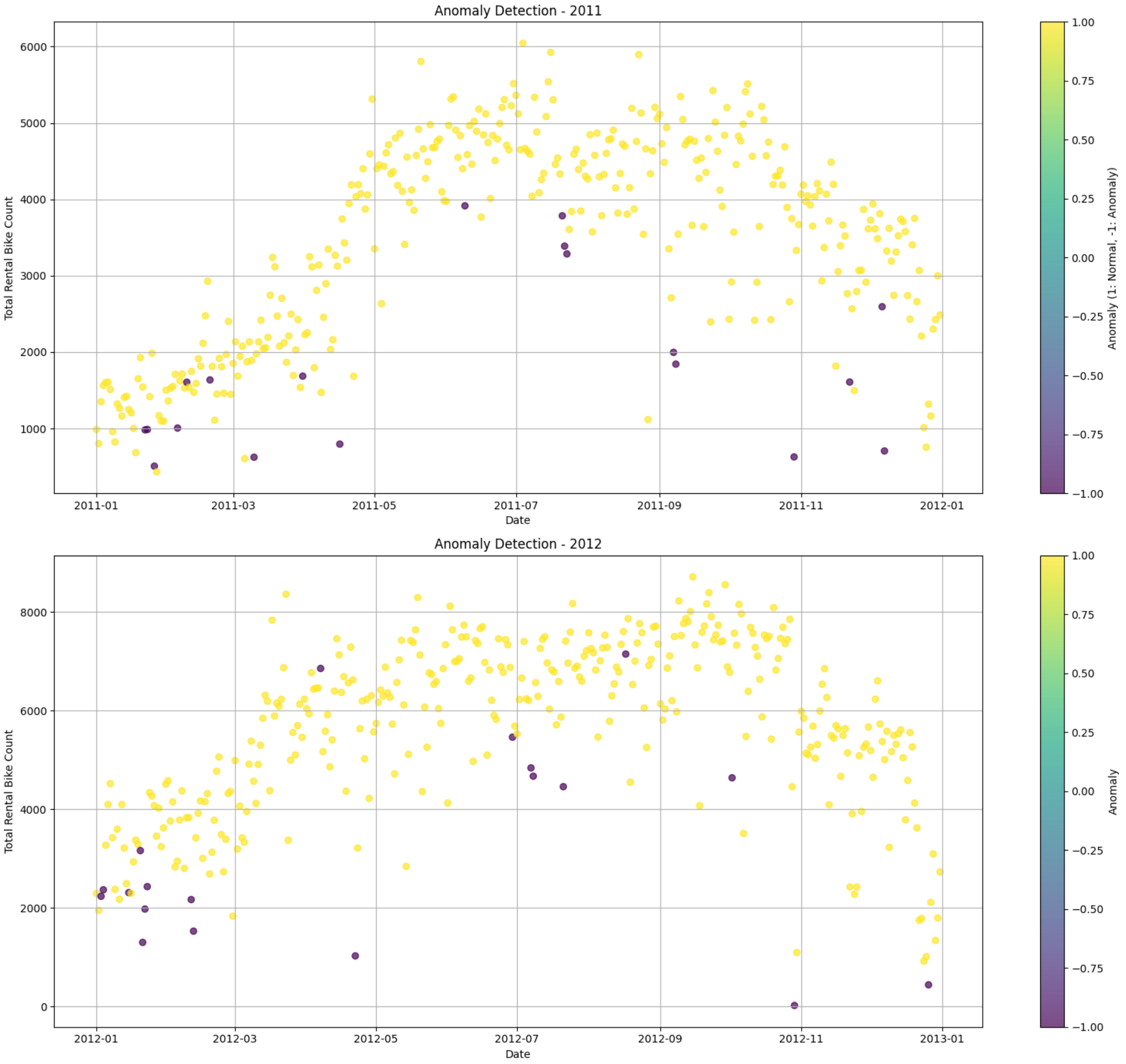

Figure 5: Year-wise anomaly detection

References

[1] Hadi Fanaee-T and Joao Gama. "Event labeling combining ensemble detectors and background knowledge". In: *Progress in Artificial Intelligence* (2013), pp. 1–15. ISSN: 2192-6352. DOI: 10.1007/s13748-013-0040-3. URL: http://dx.doi.org/10.1007/s13748-013-0040-3.