Assignment 2: Data Analysis and Visualization with Python

1. Objective: Explore Washington D.C. Bike Rental Dataset

source

The dataset can be downloaded from Kaggle here (you only need the train.csv). It provides hourly bike rental numbers in Washington D.C. for the years 2011 and 2012. The objective is to explore the effect that different weather and temporal factors have on the number of bikes rented.

2. Data Description

datetime - hourly date + timestamp

season - 1 = spring, 2 = summer, 3 = fall, 4 = winter

holiday - whether the day is considered a holiday

workingday - whether the day is neither a weekend nor holiday

weather - 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 - temperature in Celsius

atemp - "feels like" temperature in Celsius

humidity - relative humidity

windspeed - wind speed

casual - number of non-registered user rentals initiated

registered - number of registered user rentals initiated

count - number of total rentals

3. Tasks

3.1 Part I: Data Manipulation and Analysis

1. Import the dataset into a pandas dataframe. Make sure that the date column is in pandas date time format.
2. Check the data type of each column. How many rows are there in the dataset ? Does the dataset contain any missing values ?
3. Using the date column, create new columns for: year, month, day of the week and hour of the day.
4. Rename the values in the season column to spring, summer, fall and winter.
5. Calculate the total number of casual and registered bikes rented in the years 2011 and 2012.
6. Calculate the mean of the hourly total rentals count by season. Which season has the highest mean ?
7. Are more bikes rented by registered users on working or non-working days ? Does the answer differ for non-registered users ? Is the answer the same for both years ?
8. Which months in the year 2011 have the highest and the lowest total number of bikes rented ? Repeat for the year 2012.

9. Which type of weather have the highest and lowest mean of the hourly total rentals count ?
10. Calculate the correlation between the hourly total rentals count and all the numerical columns in the dataset. Which column has the highest correlation with the total rentals count ?
11. Create a new categorical column called day_period, which can take four possible values: night, morning, afternoon and evening. These values correspond to the following binning of the hour column: 0-6: night, 6-12: morning, 12-6: afternoon, 6-24:evening.
12. Generate a pivot table for the mean of the hourly total rentals count, with the index set to the day period and the column set to the working day column. What can you observe from the table ?

3.2 Part II: Data Visualisation

1. Plot the distributions of all the numerical columns in the dataset using histograms.
2. Plot the distributions of all the numerical columns in the dataset using box plots.
3. Plot the the mean of the hourly total rentals count for working and non-working days.
4. Plot the the mean of the hourly total rentals count for the different months for both years combined.
5. Plot the the mean of the hourly total rentals count for the different months for both years separately in a multi-panel figure.
6. Plot the the mean and the 95% confidence interval of the hourly total rentals count for the four different weather categories. What can you observe ?
7. Plot the the mean of the hourly total rentals count versus the hour of the day. Which hours of the day have the highest rentals count ?
8. Repeat the plot in 7 for different days of the week. What patterns can you observe ?
9. Repeat the plot in 8 for the four seasons using a multi-panel figure. What patterns can you observe ?
10. Plot the the mean and the 95% confidence interval of the hourly total rentals count versus the period of the day column, which you created in the first part of

the assignment. Which period of the day has the highest rentals count ? Does this peak period differ for working and non-working days ?

11. Plot a heatmap for the correlation matrix of the dataset numerical variables. What observations can you make ?