



Capstone Overview

In this capstone course you will apply various Data Science skills and techniques that you have learned as part of the previous courses.

You will assume the role of a Data Scientist who has recently joined an AI-powered weather data analytic company and be presented with a challenge that requires data collection, analysis, basic hypothesis testing, visualization, modelling, and dashboard to be performed on real-world datasets.

You will undertake the tasks of

- Collecting and understanding data from multiple sources
- Performing data wrangling and preparation with regular expressions and Tidyverse
- Performing exploratory data analysis with SQL and visualization using Tidyverse and ggplot2
- Performing modelling the data with linear regressions using Tidymodels
- Building an interactive dashboard using R Shiny

The project will culminate with a presentation of your data analysis report, with an executive summary for the various stakeholders in the organization. You will be assessed on both your work for the various stages in the data analysis process, as well as the final deliverable.

This project is a great opportunity to showcase your Data Science skills, and demonstrate your proficiency to potential employers.

Project Scenario

Imagine that you have just been hired by an AI-powered weather data analytics company as a data scientist.

Your first project is to analyze how weather would affect bike-sharing demand in urban areas. To complete this project, you need to first collect and process related weather and bike-sharing demand data from various sources, perform exploratory data analysis on the data, and build predictive models to predict bike-sharing demand. You will combine your results and connect them to a live dashboard displaying an interactive map and associated visualization of the current weather and the estimated bike demand.

The last assignment is creating an insightful and informative slideshow and presenting it to your peers.

Understanding the source data

Here we introduce the data sources you will be utilizing in your capstone project. Details on how to connect and/or download will be provided in the next section of this module.

Seoul Bike Sharing Demand Data Set

Data of interest

[Home](#) > [Public data](#) > [Data of interest](#)
[login](#)
[Sign Up](#)
[site map](#)


Rental bikes are available in many cities around the globe. It is important for each of these cities to provide a reliable supply of rental bikes to optimize availability and accessibility to the public at all times. Also important is minimizing the cost of these programs, in part by minimizing the number of bikes supplied in order to meet the demand. Thus, to help optimize the supply it would be helpful to be able to predict the number of bikes required each hour of the day, based on current conditions such as the weather. The Seoul Bike Sharing Demand Data Set was designed for this purpose. It contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), and the number of bikes rented per hour and date.

You will use this dataset to build a linear regression model of the number of bikes rented each hour, based on the weather.

Attribute Information

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature - Temperature in Celsius
- Humidity - unit is %
- Windspeed - unit is m/s
- Visibility - unit 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m²
- Rainfall - mm Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc (Non Functional Hours), Fun(Functional hours)

Relevant Paper and Citation Request:

1. Sathishkumar V E, Jangwoo Park, and Yongyun Cho.

Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, Vol.153, pp.353-366, March, 2020

2. Sathishkumar V E and Yongyun Cho. A rule-based model for Seoul Bike sharing demand prediction using weather data European Journal of Remote Sensing, pp. 1-18, Feb, 2020

Open Weather API Data

11:55pm, Mar 31

Seoul, KR

● 11°C

Feels like 10°C. Clear sky. Light air

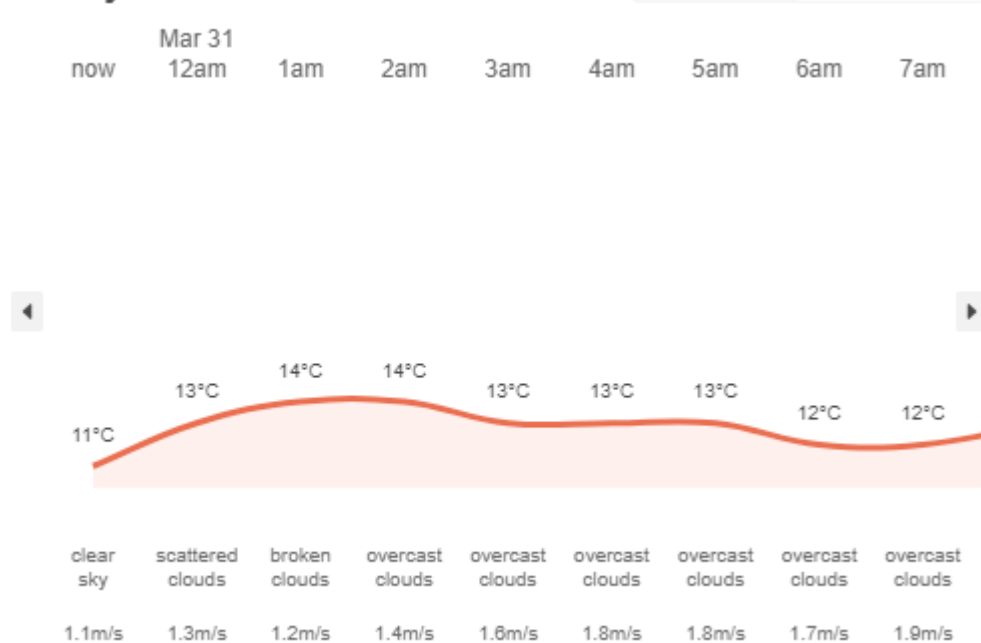
1.5m/s NW 1015hPa Humidity: 66%
Dew point: 5°C Visibility: 10.0km











Minute forecast



Hourly forecast



8-day forecast

Tue, Mar 30		21 / 9°C	scattered clouds	▼
Wed, Mar 31		21 / 12°C	overcast clouds	▼
Thu, Apr 01		20 / 11°C	light rain	▼
Fri, Apr 02		15 / 11°C	heavy intensity rain	▼
Sat, Apr 03		13 / 8°C	light rain	▼
Sun, Apr 04		17 / 7°C	clear sky	▼
Mon, Apr 05		17 / 12°C	overcast clouds	▼
Tue, Apr 06		20 / 10°C	clear sky	▼

The *Open Weather* API allows users to access current and forecasted weather data for any location including over 200,000 cities. OpenWeather collects and processes weather data from different sources such as global and local weather models, satellites, radars and a vast network of weather stations. You can access the data you will need for the project for free by registering for a Free Subscription. We will provide you with all of the details you need to sign up and to gain access to the data using HTTP request with R. The data you will be connecting to provides the weather forecast for every 3 hours over the next 5 days.

Global Bike Sharing Systems Dataset

The Global Bike Sharing Cities Dataset is an HTML table on the Wikipedia page [List of bicycle-sharing systems](https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems):
https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems.

It lists active bicycle-sharing systems around the world. Most systems listed allow users to pick up and drop off bicycles at any of the automated stations within the network.

World Cities Data

The World Cities Data contains information such as name, latitude, and longitude, about major cities around the world.

Watson Studio

For this project, you will use Watson Studio as your main development environment. Watson Studio is a component of IBM Cloud Pak for Data, is a suite of tools and a collaborative environment for data scientists, data analysts, AI and machine learning engineers, and domain experts to develop and deploy your projects.

Refer to the following link if you need any help setting up Watson Studio:

[Setup Watson Studio](#)

Next Steps

Now you should have a basic understanding about this capstone project.

In the next step of your project, you will start with collecting and connecting to these data sources.

Author(s)

[Jeff Grossman](#)

Other Contributor(s)

Yan Luo, Rav Ahuja

Changelog

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-05-01	1.1	Yan	Content updates
2021-03-30	1.0	Jeff	Created the initial version

© IBM Corporation 2021. All rights reserved.