

## Case Study (Advanced)

### BACKGROUND

BeachBoys BikeShare is a bike share service provider where users can take and return bikes at any of the 70 stations on their network. The company wants to leverage their data to better understand and, hopefully, optimize their operations. BikeShare has decided to start by harnessing analytics to enhance operations in the logistics department, by improving the redistribution of bikes between stations to meet demand, and ensuring that there are bikes and return docks available when and where users need them.

As a key step towards tackling this challenge, management has tasked you to develop a model capable of predicting the net rate of bike renting for a given station, which is defined as the number of bikes returned to, minus the number of bikes taken from, the given station in a given hour. In other words, your model should enable BikeShare's logistics team to make the statement - "In the next hour, the quantity of bikes at station A will change by X".

In addition, management would also like you to:

- Help them understand the factors that affect bike rental, which could inform future decisions on where to locate BikeShare's stations.
- Help them conceptualize how your prediction may be used to improve the redistribution of bikes within the network.
- Highlight any assumptions or drawbacks of the analysis, if any, and suggest how they may be verified or addressed in the future.

### ASSIGNMENT

Explore, transform, and visualize the given data as appropriate, before using it to train and evaluate an appropriate ML model for the problem. Address the issues highlighted by management as described above, albeit in a less in-depth manner. Finally, articulate your findings and recommendations in a concise, coherent, and professional manner, making reference to any earlier results or diagrams as appropriate to support your conclusions.

Please use Python to complete this task, using any libraries you might deem necessary for your analysis, e.g., pandas, sklearn, etc. Detail your code, analysis findings, and recommendations clearly in a reproducible Jupyter notebook with appropriate comments and documentation, so that an individual viewing the notebook will be able to follow through your steps and understand the reasoning involved and inferences made.

### DELIVERABLES

You should upload the following deliverables in a .zip file:

- A Jupyter notebook detailing your analysis and findings for this project,

- A PDF-ed copy of the Jupyter notebook above, which should not exceed 30 pages,
- The datasets used in your analysis, which should be loaded into your notebook,
- Additional files relevant to your analysis, which should be described in your notebook.

Finally, please prepare presentation slides for the group presentation (15-20 mins). All team members need to present in English.

## THE DATA

The company has collected information on the stations, trips taken, and on weather conditions in each of the cities from September 2014 to August 2015. You can find the [data here - bikes\\_data.zip](#) (3.1 MB). Below, you will also find detailed information on all the fields available in the dataset. The way you include this information in your model is up to you and should be clearly justified and documented in your report. You are free to use any other data sources provided you specify a link to this information in your report.

### Station Data

- Id: station ID number
- Name: name of station
- Lat: latitude
- Long: longitude
- Dock Count: number of total docks at station
- City: one of San Francisco, Redwood City, Palo Alto, Mountain View, or San Jose

Please note that during the period covered by the dataset, several stations were moved. Stations 23, 25, 49, 69, and 72 became respectively stations 85, 86, 87, 88, 89 (which in turn became 90 after a second move).

### Trip Data

- Trip Id: numeric ID of bike trip
- Duration: time of trip in seconds
- Start Date: start date of trip with date and time, in Pacific Standard Time
- Start Station: station id of start station
- Start Terminal: numeric reference for start station
- End Date: end date of trip with date and time, in Pacific Standard Time
- End Station: station id for end station
- Subscription Type: Subscriber (annual or 30-day member) or Customer (24hour or 3-day member)

## Weather Data

- Date: day for which the weather is being reported
- Temperature (day min, mean and max): in F
- Dew point (day min, mean and max): Temperature in F below which dew can form
- Humidity (day min, mean and max): in %
- Pressure (day min, mean and max): Atmospheric pressure at sea level in inches of mercury
- Visibility (day min, mean and max): distance in miles
- Wind Speed (day max and mean): in mph
- Max Gust Speed: in mph
- Precipitation: total amount of precipitations in inches
- Cloud Cover: scale of 0 (clear) to 100 (totally covered)
- Events: Special meteorological events
- Wind Direction: in degrees
- Zip: area code for San Francisco (94107), Redwood City (94063), Palo Alto (94301), Mountain View (94041), and San Jose (95113)