Technical Challenge: Data Science

In this technical challenge we would like you make future usage predictions for a bike sharing program. Attached you will find a dataset that contains the historical daily usage of rental bikes between the years 2011 and 2012 with corresponding weather and seasonal information.

Column Descriptions

For the attached dataset please find the following column description below

```
- 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)
- holiday : whether day is holiday or not
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
+ weathersit :
- 1: Clear, Few clouds, 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. The values are derived via (t-
t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- atemp: Normalized feeling temperature in Celsius. The values are derived
via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: coundata t of registered users
- cnt: count of total rental bikes including both casual and registered
```

Data source: http://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

Task Description

We would like you to predict the total number of bikes shared daily for the same program for the future years. Below you will find a series of questions and tasks that will help us better understand how you think about, evaluate, and solve this challenge.

1. Do you think the included data will be useful in training a predictive model for this purpose? Which variables would you expect to have some effect on the number of shared bikes on a given day? Explain your reasoning.

2. Which feature represents the dependent variable and which ones would you consider independent variables for this problem? Which columns can you discard since they do not provide valuable information? Would you expect high correlation between some of the variables? What should we do with these correlated variables?

3. Train a linear regression model on the independent variable using your chosen dependent features. (Please use Python or R)

    a. How do you preprocess your data? Are there missing values? What would you do if you had missing values? When do you think standardization would be necessary?

    b. What about categorical variables? When does label encoding make sense?

    c. How many dimensions do you have after preprocessing? Is it too many? Why? How would you reduce the number of dimensions when you look at your training results?

    d. Which feature(s) do you think have the most effect on the dependent variable from your training results?

    e. How do you assess the success of your obtained model? Report the Adjusted R-squared, F-statistic and the MSE (mean squared error) error on your training and test data.

    f. Plot your predictions against the real values (the x-axis can be the date or the index). What do you think about the results? Is the fit good? How would you recommend finding a better model?

4. Consider this data was in an SQL table with features as columns. Write the SQL statement to get the average daily number of shared bikes monthly (calculate also variance of daily shared bikes for each month in two years). Do the same with pandas. Plot the distribution of average daily number of shared bikes against month/year (x-axis is the month/year).

5. (OPTIONAL) Consider you had another file with the following fields:

    ```
    - dteday : date
    - traffic : 0 for low 1 for medium 2 for high
    ```

    Assume you might not have data for all days for this table. How would you merge this new data with the existing data? Explain with pandas and SQL. How would you preprocess this new feature before training your linear model?

Requirements

- There is no deadline to complete this challenge, but it should not take more than ***four hours*** to finish. We’re looking for structure, not complexity.

Evaluation Criteria

- Critical thinking
- Data analysis and interpretation
- Visualization methodology
- Processing and handling of data
- Code structure and quality
- Documentation clarity and structure

Deliverables

Please send your report, your code and any of your documentation packaged in a single file (.zip, .tar, etc.) via e-mail attachment to your interviewer.

Have Fun, Happy Coding!