Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Data about the point in time (weekday, date, season) and the weather (temperature, humidity and windspeed) explain the number of rentals quite well (81.6%). The meaningful categorical values however here are only the seasons. The year is only technically a categorical values (2018 or 2019) but since the year can not be repeated they don't inform the prediction for 2020 onwards.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   The information if the first column of dummy variables is true can be inferred from the others (all False) therefore it can be dropped without loosing information. This reduces the number of variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   The day/date

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   With statsmodel's descriptions, a plot of the error terms and by making predictions on the test set to see if the results are plausible.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   Humidity, temperature and weather situation

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   The linear regression algorithm fits a line (linear) to a distribution iteratively (regression). A simple linear formula $Y = \beta_0 + \beta_1 X$ is adjusted to better fit the distribution by reducing the sum of the distances of all points to the linear regression line.

2. Explain the Anscombe's quartet in detail. (3 marks)

   Anscombe's quartet is a collection of datasets (x, y) that are equal in mean, standard deviation, and regression line, but look very different when plotted. It is used to illustrate the importance of looking at a set of data visually and not only relying on basic statistics.

3. What is Pearson's R? (3 marks)

   The Pearson correlation coefficient (r) is a common way of measuring correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It can be used to assess a linear correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Scaling deals with the "level" and unit of a variable. Having the unit change from 1000s to single singles would otherwise change the term for a linear regression. We standardize/normalize a variable to solve this problem. When we standardize we make the variable have a mean of zero, and a standard deviation of 1. When normalizing we make the variables minimum value be 0, and the highest value 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   VIF approaches infinity when the correlations get bigger. With perfect correlation VIF becomes infinite (actually undefined) as we divide by zero ($VIF = 1/(1-R^2)$ while $R = 1$)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

   With a Q-Q plot one can check if the residuals of a linear regression model are normally distributed.