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)

From the box plots: Holidays have a negative effect on the rentals with higher uncertainty (higher variance). Mist has slight and rain has significant negative effect that coincides with domain specific knowledge. Data shows that in spring the daily rentals are significantly lower.

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

To represent a categorical variable with with N independent states, N-1 dummy variables needed. Without dropping (using N) adds redundancy and multicollinearity that should be avoid.

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

temp, atemp, (they also have very high correlation between each other)

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

Checking if the residuals are normally distributed. Try to find patterns in the residuals visually, if possible, to visualize.

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

From coefficients: temp, rain, yr

General Subjective Questions

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

- Read data
- Check for missing or invalid data, apply var or row removal or data imputation. Convert data if needed.
- Identify numeric and categorical variables
- Make derived variables if applicable.
- Plot numeric variables to visually check for linear relationship and correlation between target variable and independent variables.
- Plot categorical variables (i.e. boxplot) to visually check relationship with target variable
- Calculate basic statistics (i.e. correlation matrix) to numerically check relationship (interpret numerical and graphical results together)
- Convert categorical variables to dummy variables (N independent state, N-1 dummy var)
- Scale variables (normalize or standardise). (recommended)
- Select the best n variables for the model describing the most variance in the data set with the least number of variables.
  Iterative process: Manual (Forward, Backward), Auto (i.e. RFI), Mixed (coarse with auto, fine

tuning by manually adding and removing variables and checking the results)
Indicators: P value for a variable to check significance, VIF to detect multicollinearity, R2 to assess how the model describes the variance in the data (is good enough?).

- Train the model with the selected set of variables
- Residual analysis: check if the residuals are normally distributed, look for patterns in residuals (should not be any)
- Model evaluation: make the prediction with the trained model on the test set. Check R2 on the test set. Should not be too large difference between R2 train and test R2. Mean square error or other metrics can also be used to evaluate the trained model.

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

Anscombe's quartet comprises of 4 rather different data sets (11 points each) with almost identical descriptive statistics. (Mean X, Y, Sample variance X, Y, Correlation (X,Y), Linear Regression coefficients, R2)

The datasets were created by Francis-Anscombe to demonstrate the limits of simple descriptive statistics and encourages statisticians to also look at the data graphically. Visual checking can help outlier detection and identifying cases where there is no linear relationship between the data sets, therefore linear methods are not applicable.

- First data set shows linear relationship and correlation between two variables. Dependent var also normally distributed.
- Second shows a non-linear relationship. Correlation coefficient and linear regression are not applicable for the data and lead to faulty results.
- Third shows how even 1 outlier can influence the LR model and the correlation coefficient.
- Fourth shows how one high-leverage point (outlier) can generate a high correlation coefficient where there is no correlation between X and Y.

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

Correlation coefficient. Measures the linear correlation between two variables. Normalized covariance.

R(X,Y) = covariance(X,Y) / (stdev(X)*stdev(Y))

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

Data can highly vary in magnitude, unit, range, etc. Keeping the original scale will result in a valid model but the coefficients will not be easily interpretable or at all interpretable. Scaling is a pre-processing step to normalize the data to a more desired range or scale that result in a better interpretation of coefficients. Model training usually uses some unconstrained optimization algorithm to minimize the cost function, with scaling the algorithm converges faster to a local minimum.

- Normalized scaling (Min-Max): Converts data into the range of 0 and 1.
  x=(x-min(x)) / (max(x)-min(x))
- Standardized scaling: Converts data into a standard normal distribution having mean=0 and stdev=1.
  X=(x-mean(x)) / stdev(x)

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

(3 marks)

Perfect correlation (R(X,Y) =1) between two variables lead to infinite VIF.

$VIFi = 1 / (1-R_i^2)$

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

(3 marks)

Quantile -Quantile plot is a technique that helps visually identifying if two set of data came from a population with the same distribution. Or if one set of data is verified to came from or generated from a theoretical (i.e. Normal) distribution the plot helps to assess if the other set of data comes from the same distribution.

If the two data set come from the same distribution the plotted points lie close to a theoretical straight 45 degree line (from x-axis).

Points away from the theoretical line indicates different distribution.