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)

1. **"Month"**: There is a strong correlation between the monthly rental (the dependent variable) and 'mnth' variables. The monthly mean rental continuously increases from January, reaching a plateau around June, and starts declining again somewhere around October, November. The overall spread of the data is the biggest in March and October. Two months where the weather is less predictable. It all suggests that this effect is caused by weather conditions.
2. **"Season"** shows the same pattern as "Month" (the mean target value increases from winter through spring, reaches the top in summer and declines again with the fall coming). No wonder as it is an aggregated value (months are mapped to seasons)
3. This becomes more clear when we look at the **"weather situation"** variable. The weather situation has a strong effect on the bike rentals. The better the weather the higher the bike rentals are.
4. **"Workingday"** has no no strong impact on the rental numbers. That was a bit counterintuitive as I expected a larger impact caused by the fact that people behave differently when they work and when they don't. It would be interesting to see why there is no strong, visible impact. Is it because the bikes are mainly not by locals for commuting but let's say by tourists? Further investigation would be interesting on this.
5. The "**Weekday**" variable shows the same pattern (or to be more specific the lack of the pattern) as the "Workingday". Again, this is expected as Weekdays are mapped to working days and non-working days.
6. **"Holiday"** variable shows that people tend to rent less bikes on holidays. Also, the overall spread of the data is smaller than it is for non-holidays. However, we only have very few observations in the dataset where "holiday == 1". Therefore, it is hard to conclude anything from it as the dataset is pretty much unbalanced from this perspective.
7. **"Year"** has a strong impact on the target variable. The company gained traction from year 0 to year 1 and the numbers are shifted to the right. Also, the shape of the distribution is different for the two years suggesting a structural change. Instead of simply resulting in higher numbers, the distribution is slightly different for year 1. Is it because for year 0 most of the users were "early adopters", and for year 1 "fast followers" stepped in with slightly different behaviour? Perhaps this is an explanation. Or did they expand the number of available bikes somewhere around the year?Another interesting fact is that the curve is left skewed. Why is that? What causes this skewness? Is there an upper bound? Is there a limit perhaps in the number of bicycles that are available for rent? Well, certainly there is. So the question here is if it is responsible for the skewness? My working theory is that there is a physical limit of bikes available for rent. This limit can't be exceeded. It would be amazing to analyse some "utilisation" numbers to verify this theory. If it is the case, it is very likely that increasing the number of bikes might generate some extra income.

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

If we don't drop the first dummy variable, it means that the number of variables will be equal to the number of categories. It has two important impacts:

- The model becomes unnecessarily complex as we could perfectly describe n categories by n-1 variables. (Unnecessary complexity might cause overfitting in certain cases)
- In case we don't drop the first dummy variable and we end up using n variables for n categories, any of those n variables will be a perfect linear function of the remaining n-1 variables. In other words, we introduce multicollinearity in the model.

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

"temp" and perhaps "atemp" so the temperature.

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

Distribution of error terms ("normality")
The **histogram** of the error terms clearly showed that distribution of the error terms approaches normal with mean 0. It means that the "Normality" criteria of the error terms is fulfilled.

Error terms are independent of each other
The **error terms vs. observation order** scatterplot showed no real pattern. From this one can conclude that the error terms are independent of each other

Homoscedasticity
The **predicted y vs. error terms** scatter plot shall show no difference in the variance of the error term. Well, in our case the variance of the error term is not quite constant. The variance of the error term is a bit higher for higher predicted values. Hence, we observe slight heteroscedasticity.

Multicollinearity
I calculated the VIF values for each variable. Based on the result one can conclude that no independent variable is a perfect linear combination of other independent variables hence there is no multicollinearity in the model.

Linearity
The regression model is linear in the coefficients and the error term.
The model has high R squared value which doesn't drop significantly when calculating it on "unseen" test data.
The overall model is significant (P value of the T statistics)
All the coefficients are significant (p values are below 0.05 for each). All in all, the model correctly fits the linear pattern.

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

In case of my final model these features (in descending order of the absolute value of coeffs) are:

- atemp (feeling temperature)
- heavy_precip (heavy rain / snow)
- Yr (indicates first or second years for the observation)

General Subjective Questions

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

Linear regression is a machine learning algorithm. It is a supervised learning where **our goal is to learn a function that best approximates the relation between the independent variables and the target variable** based on a dataset of prior observations.

As almost every other machine learning model, linear regression is built up by the following meta structure:

1. Vectorising the input dataset (we skip this part for the current explanation)
2. Choosing a model that describes the relation between the target and feature variables best
3. Choosing a cost function that expresses a real 'domain related' cost caused by the prediction errors of the model
4. Choosing an optimisation method to find the minimum of the cost function by fine-tuning the parameters of the model

Model

As the name suggests it, the linear regression assumes a linear relation between the independent variables and the target variable. The linear relation is what it can model. In case of one independent variable, it can be visualised as a single line ($y = ax + b$), in case of multiple independent variables this linear relationship is represented by a hyperplane ($y = a1*x1 + a2*x2\ldots\ \ldots + an*xn + b$) in the multi-dimensional Euclidean space.

In case of simple linear regression, the coefficient of the independent variable tells by how much the value of the target variable changes in case the value of the independent variable moves by one unit. In case of multiple linear regression (more than one independent variables) the coefficient of variable n shows by how much the value of the target variable changes by moving the value of variable n by one unit and keeping the rest of variables unchanged.

Cost function

The cost function for linear regression usually captures the error the model makes when predicting. In case of linear regression, the cost function is always some kind of distance function between predicted and observed points. Simply saying it measures the sum of the differences between the predicted and the observed target variable for each observed

datapoints. The problem with measuring the pure difference is that the error can be both positive and negative. The individual errors could net each other out when summing up resulting in a low total error even though the model makes huge errors for each point.
To avoid this situation it is logical to use the absolute value of the difference for every datapoint abs(y - y_pred). It eliminates the problem of having both positive and negative errors, however it doesn't differentiate between small and large errors. Therefore the commonly used cost function is the Sum of squared errors (SUM((y - y_pred)**2) for each datapoint. This approach does not only eliminate the problem of different signs but also penalises larger errors more hence it is an appropriate cost function.

Optimization
The last question remains here is how to find the optimal fit, in other words how to find the minimum value of the cost function. Mathematically there are multiple solution for solving this problem. We can approach it as a matrix algebra task, or we can approach it with a more iterative approach called gradient descent. In case of gradient descent, we calculate the derivative of the cost function. Depending on the absolute value and the sign of the result we change the model parameters and calculate the derivative again until the value converges to the minimum.

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

Anscombe's quartet is a collection of four data sets. Each of these datasets consist of 11 (x,y) pairs. These 4 sets are interesting as these they have nearly identical descriptive statistics but plotting them shows a huge difference in distribution.

- Mean and sample variance of x

are identical for all 4 sets.

- Mean and sample variance of y
- Correlation between x and y
- Intercept and coefficient of x of the regression line
- R squared of the regression

are nearly identical (identical to 2 or 3 decimals)

These sets demonstrate the importance of data visualisation as they clearly show that patterns which can easily be observed by visualisation might remain uncovered by checking the numerical descriptive statistics only. Also emphasises the impact of outliers and "other influential observations on the statistical properties".

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

Pearson's R or Pearson's correlation coefficient is a bivariate analysis. It measures the strength and the direction of correlation between two variables.

In statistics we quite often want to measure how two variables are related to one another. To measure this relationship we need to have (x,y) pairs as datapoints. If for most of our (x,y) value pairs a relatively low x value comes with a relatively low y value, and equally a relatively high x value comes with a relatively high y value, this relationship is positive.
Similarly if a relatively high x value comes with a relatively low y value, while relatively low x value comes with a relatively high y value, the relationship is negative.

One convenient way to quantify this relationship is the **covariance**: Cov(X,Y) = E((x-xmean)*(y-ymean))

The intuition behind the formula is that if x is relatively small (i.e. smaller than its mean), x-xmean will be negative. When the relationship is positive y tends to be relatively small with relatively small x, meaning that y-ymean will also be negative, thus the product of these two values will be positive and so on.
So the covariance is an intuitive way for expressing negative, positive or 0 relationship. But there are some downsides of this formula. It can take any value from negative infinity to positive infinity. It means that even though it is able to express the direction of the relationship between two variables, it doesn't reflect the strength of this relationship and the absolute value of covariance is very hard to interpret.

This is where **Pearson's correlation coefficient** comes into picture. It divides the value of the Covariance by the product of the square root of variance x and square root of variance y

Cov(X,Y)/(sqrt(Var(x))*sqrt(var(y)))

The intuition behind the formula is the following. The value of the Covariance depends on the direction of the relation between x and y, how far the data points are spread around mean, and the scale of the variables. In the formula of the correlation coefficient the denominator squeezes the value of the covariance between -1 and 1. It in fact it 'normalises' it, eliminating the impact of different scale and different spread around mean.

As a result, the correlation coefficient takes the value from -1 to 1. When the correlation coefficient is 1, x and y have a perfect positive correlation. In other words, when value of x or y moves the other value moves in the same direction, proportionally as well. In case of -1 the correlation is perfect negative correlation. In case of 0 there is no correlation between the two variables. Moving from - 1 towards 0 or from 1 towards 0 indicates a weaker and weaker correlation.

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

Feature scaling is an exercise usually during the data preparation phase when we transform the numeric variables so that all the numeric variables can be nearly on the same scale.
There are several reasons to run scaling.

1. It is essential to do it for models that optimise by calculating distances. When the scales are very different, the variable with larger scale "starts dominating" the optimisation process and

the model. Having the variables on the same (or near) scale results in better distance calculations, where the different scale doesn't have an impact on the overall distance.
2. In case of gradient descent, the convergence happens much faster if the variables are on the same scale since we have a more "spherical" problem to solve.
3. The comparison of final coefficients is easier if the variables are on the same scale. In this case one can conclude that the variable with higher absolute coefficient value has greater impact on the target variable.

Normalised scaling or also called Min-Max scaling is a scaling method where we squeeze the values of the variable between 0 and 1

Standardisation also known as z-score scaling rescales the data the way that it has a mean of 0 and a standard deviation of 1

In case of min-max scaling our variables have a fixed range ([0,1]). It can suppress the impact of outliers. In case of normalised scaling the range is not fixed, but the distribution is centered around 0.

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

VIFi = 1/(1-R2i). Infinite VIF happens when R2 is infinitely close to 1. It means that the independent variable is a perfect linear function of other independent variables. There is a perfect correlation between the respective variable and some other variables.

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

Q-Q plot or quantile to quantile plot is a graphical way to compare and analyse two distributions by plotting their quantiles against each other.

If the two distributions are perfectly identical, the plotted points will lay on the x=y line.
In practice it is quite often used to compare not only two data samples but also the distribution of a sample to a theoretical distribution.

One can answer the following questions using this plot:
- Do the two data sets have common distribution?
- Do the two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?