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)

Looking at the variation of box plots of categorical variables from the dataset:
* Season: clear seasonal pattern with highest rental in summer and fall
* Year: 2018 (yr=1) shows a higher rental, proving a growing adoption of the service
* Month: a sinusoidal (wave-like) pattern with the peak in summer and a drop in spring
* Holiday: slightly wider spread on holidays (holiday = 1). It seems that holidays have more variability on bike demands while non-holidays have more consistent rental patterns. This can be explained by the different types of holidays, events and activities impacting the bike usage behavior.
* Weekday: relatively consistent over the weekdays suggest a steady commuter usage. Variability during weekends.
* Workingday: very similar patterns suggest that the working schedule may not influence the bike usage.
* Weather condition: shows a highest rental demand during clear weather while light rain shows a lowest rental. There is no data for the "heavy rain" category, so users will probably avoid cycling in heavy rain.

We can infer that season, weather, yearly growth and month are the strongest categorical predictors of bike rentals while others have less impact. Holiday, weekday and working day show some influence, however mainly in terms of variability rather than the overall rental demand

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

It's essential to use drop_first=True during dummy variable creation to avoid the dummy variable trap, which creates multicollinearity. A categorical variable with n categories, (n-1) dummy variables are enough as the nth category can be inferred from the others.

A dummy variable trap occurs when we include all n dummy variables in the model, creating multicollinearity because the value of one dummy variable can be predicted from the others. For example, a "Gender" variable with "Male" and "Female" category: we just need (n-1) = 1 category as if Male = 1, then Male = 0 can be inferred as Female.

By setting “drop_first = True” in the creation process, the reference category is omitted, leaving only the meaningful categories for modelling. This approach is significant to prevent multicollinearity, simplifying the model, maintaining the model accurate, reliable and stable

However, when using “drop_first=True” in the function pd.get_dummies(), it will automatically omit the first category. For example, if the categorical variable has A, B, C and D, then category A will be dropped. In practice, we may need to choose a meaningful reference category or select a duplicated category to drop, in that case, we may want to drop selectively instead of using “drop_first”.

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

Temp and atemp show a strong positive correlation with bike usage while humidity and wind speed have a weak correlation due to a very scattered pattern

This suggests that we can apply the linear regression to predict the bike users as the linear relationship exists, potentially temp/atemp are strong predictors with the highest correlation with the target variable. In terms of business, more bike rentals occur in warmer temperatures.

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

The key assumptions of linear regression that we need to validate: linearity, homoscedasticity, normality of residuals and independence of errors.

We can perform several validation approaches

1. Check the linearity assumption and homoscedasticity using a scatter plot of the predicted values and residuals: the plot should not show any patterns and trends.
    - Linearity: the plot should not show any systematic patterns or trends. Any visible patterns such as a curve suggest a non-linear relationship, suggesting that using a linear model is not suitable, we may have to consider another model.
    - Homoscedasticity: the spread of residuals should be consistent across. If it increases or decreases as the predicted value changes, it violates the assumption of constant variance.
2. Normality of residuals: check the error assumption using the distribution plot of the residuals: residuals should show a normal distribution and randomly around zero.

If the errors have deviation at tails, we can use Q-Q plot to identify patterns of deviation. In the Q-Q plot, the points should lie approximately along a straight line. Deviation from the line suggests non-normality.
For example, sales data may have some extreme values, creating heavy tails of residuals which the model may not be able to predict well. We can consider log transformation to reduce the impact of outliers.

3. Independence of error: Check the statistical independence of the errors using statistical test such as
    - Durbin-Watson test: value around ≈ 2.0 indicates that the errors are independent. Values close to 0 or 4 suggest positive or negative autocorrelation
    - Breusch-Godfrey test: a p value > 0.05 desired to indicate that there is no significant autocorrelation

    Autocorrelation in the residuals indicates that the errors are not independent, causing poor prediction. For example, in time series data, this month sales may be influenced by last month sales, indicating a dependency.

Without checking the model assumption, we can not have a reliable model, subsequently cannot make our inference.

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

The top 3 variables contributing most significantly to the model are:

1. yr (coefficient: 2033.45)
    - Highest absolute coefficient
    - Very significant (t=26.9, p<0.000)

    - Strongest positive impact with the increase in bike demand. The coefficient indicates that for each year, the demand of bikes increases by around 2033 units, assuming all other factors remain the same.

    This suggests a growing trend in bike usage over the years may be due to the city development or improving in brand awareness or services.

2. temp (coefficient: 785.01)
    - High positive coefficient
    - Highly significant (t=14.8, p<0.000)

- Strong positive impact on bike demand. For a unit increase in temperature, bike demand will increase about 785 units, assuming all other factors remain the same.

This suggests that warmer weather will boost the demand for using bikes, likely due to the better riding conditions or increasing outdoor activities.

3. season_spring (coefficient: -1316.8)
    - Large negative coefficient
    - Highly significant (t=-11.8, p<0.000)

- Strong negative impact on bike demand. During spring, the demand for bike decrease around 1316 units compared to the reference category (the dropped season “autumn”)

This suggests a significant reduction in demand during spring probably due to weather conditions or certain patterns of user behavior different from other seasons.

General Subjective Questions

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

Linear regression models the relationship between a dependent variable (Y) and independent variables (X). The goal of linear regression is to find the best linear equation that describes the relationship between (Y) and (X). If a single variable X, it’s called single linear regression. If there are multiple variables X, it’s called multiple linear regression. The relationship equation can be generalized as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$

Where

- β_0 is intercept, $\beta_1 \ldots \beta_n$ are coefficients showing impact of each feature (X variable)
- ε represents error term/residuals

The algorithm finds optimal coefficients $\beta_0, \beta_1 \ldots \beta_n$ that minimize prediction errors defined by

$$\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$$

where $\widehat{Y_i}$ is the predicted value of Y for the i-th observation

This is also called the sum of squared residuals or the cost function of the model. To minimise the cost function, we use a method called **Ordinary Least Squares**

Linear regression is validated based on assumption that

- Linearity and additivity of the relationship between dependent and independent variables: The relationship between X and Y should be linear.
- Statistical independence of the errors: The residuals should be independent of each other or no autocorrelation. If not, it indicates that errors are dependent from each other, creating a poor accuracy for the model
- Error terms are normally distributed: The residuals should be normally distributed, centering around 0. If not, the model will mispredict by a certain amount, indicating an improper relationship between X and Y.
- Homoscedasticity: The variance of the residuals should be constant across. If not even we can fit a line through the data, we cannot make inferences about the model.
- No Multicollinearity (for multi linear regression): The independent variables should not be highly correlated with each other. With multicollinearity, the model's performance will become unstable with high errors. A small change in the data can lead to a large change in the coefficient estimates. Meanwhile we cannot interpret the impact of the target variable as the effect of the variable is not isolated. The model prediction power will be reduced and we will get an overfitting model in the training data.

Linear regression is widely used in forecasting economic and marketing indicators or modeling relationships between pricing and other market factors. It's easy to interpret the impact of each feature on the target variable using the coefficients. However, it also has several limitations and considerations:

1. Linear regression is based on several assumptions mentioned above, if violated, leading to biased and inaccurate prediction power. Therefore, it's very important to always validate the model before making any inferences.
2. Linear regression sensitive to outliers and distorting the prediction result. Outliers increase the variance of the coefficient estimates, pulling the regression line towards them, leading to bias in prediction.
3. Linear regression cannot capture nonlinear relationships which are more common in real life. For example, the relationship between the age of a car and its resale value is nonlinear as cars get older, their values decrease at a certain rate.
4. Linear regression can be applied to predict a continuous numerical outcome such as sales amount, stock prices and not suitable for categorical outcomes.

5. Linear regression cannot predict out of the trained range. As it's based on the data used to train them to make predictions, it assumes the relationship remains constant in the future. However it will not account for potential changes outside the data range, leading to potentially inaccurate predictions beyond the observed data.
6. Linear regression can suffer from overfitting if we have too many variables and sample size is small.

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

The Anscombe's quartet created by Francis Anscombe consists of four datasets that have identical summary statistics (mean, variance, correlation, regression line), but completely different distributions when plotted.

All datasets have the same linear regression line Y = 3 + 0.5X, however, the underlying structures are completely different:

- Dataset 1: when plotting, it forms a linear pattern closely to the line
- Dataset 2: it's a curved pattern (non-linear relationship) with outliers significantly affect the regression line
- Dataset 3: a linear relationship with a single outlier pulling the line away from the rest of the data
- Dataset 4: the plot shows a horizontal line

We can learn from the Anscombe's quartet that

- Visualization of data is very crucial to evaluate the dataset for modeling. Even with same statistics, visual data helps to reveal differences and distribution of data for accurate modeling
- Limitations of summary statistics: statistics won't provide a full picture, sometimes mislead our interpretation.
- Outliers and non-linearity: the impact of outliers and non-linear relationships that affect linear regression may not be realized from only summary statistics.

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

Pearson's R is a measure of linear correlation between two variables. Formula:

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- σY - standard deviation of Y
- σX - standard deviation of X
- μX - the mean of X
- μY - the mean of Y
- E is the expectation.

It measures strength and direction of linear relationship, ranging from -1 (perfect negative) to +1 (perfect positive). Zero indicates no linear relationship.

It's used in statistical analysis and feature selection. By checking Pearson's R of features with the target variable, we can select features with high correlation, suggesting a strong predictor. As checking the Pearson's R of all feature variables with correlation matrix, it helps to remove multicollinearity

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

Scaling is the process of transforming numerical features to a similar range. It's performed because scaling helps to make features comparable and improve model performance.

<u>Min-Max Scaling or Normalized scaling:</u>

getting values between 0 and 1 (normalizing)

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

This might be useful in some cases where all parameters need to have the same positive scale. Guarantees all features will have the exact same scale but does not handle outliers well.

Linear scaling is a good choice when all of the following conditions are met:

- The lower and upper bounds of your data don't change much over time.
- The feature contains few or no outliers, and those outliers aren't extreme.
- The feature is approximately uniformly distributed across its range. That is, a histogram would show roughly even bars for most values.

<u>Standardized Scaling</u>

subtracting mean and dividing by σ such that the values will be centered around 0 with std = 1

$$x = \frac{x - mean(x)}{sd(x)}$$

it is recommended only when "populations that are normally distributed". Handles outliers, but does not produce normalized data with the exact same scale

- Z-score uses standard deviation, which is less sensitive to outliers than range
- Resulting values distribute around 0, with most falling between -3 and +3
- Preserves relative differences between normal values better

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

There are 2 scenarios where the $R^2=1$ so $1-R^2=0$ causing the VIP to be infinite. They are:

- **Perfect multicollinearity** exists when one or more variables in the model are perfectly correlated with each other.
- **One variable is an exact linear combination of others**, in other words, one independent variable is a linear function of the other independent variables with no error. For example: $X_3 = 2X_1 + 3X_2$

Why does this happen?

- **Dummy variable trap:** including all levels of a categorical variable in the model leads to perfect multicollinearity. To avoid this, we need to always omit 1 reference category by dropping it manually or using drop_first=True
- **Including derived/redundant features:** for example, we may include both the total sales and the individual sales of a product in the model leading to perfect multicollinearity.
- **Perfect correlation between predictors:** this can happen when the same variable with different scales is included. For example, we may add in the temperature in both Celsius X_1 and Fahrenheit ($X_2 = 1.9X_1 + 32$)

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 visualization tool to assess the normal distribution of a dataset.

- Plots theoretical quantiles (normal distribution) vs actual quantiles of the actual dataset.
- Straight diagonal line indicates normal distribution

Q-Q plot is Important for checking assumptions of linear regression models. It's performed with the residuals of a linear regression model to validate the model normality assumption. It helps to ensure that the model's inferences are reliable in regression analysis. If there are deviations from the straight diagonal line, indicating non-normality. We may have to consider corrective actions such as data transformation or model adjustments.