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)

Ans:

- Seasons: Each season appears to have a different impact on bike rentals, fall season seems to have a relatively higher positive impact on bike rentals compared to summer, spring and winter. This suggests that the season has an effect on bike rental counts, with more rentals occurring during the fall season.
- Year: In 2018 the count of bike rentals is very high compared to the 2019. This indicates a positive trend in bike rentals over the years.
- Month: September and October month has seen a rise on bikes booked compared to the other months. These months have a positive effect on the dependent variable.
- Holiday: holidays do have an effect on bike rental counts, with fewer rentals occurring during holiday periods.
- Weekday: Day of the week influences bike rental patterns, with Wednesday showing the highest bikes rental day.
- Workingday: Working days have a higher median count of bike rentals compared to weekends/holidays, suggesting that working days contribute to increased bike rental demand.
- Weathersit: Weather conditions play a crucial role in bike rental patterns. Clear weather is most favorable, followed by cloudy conditions, while heavy rain/snow has the least favorable impact on bike rentals.

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

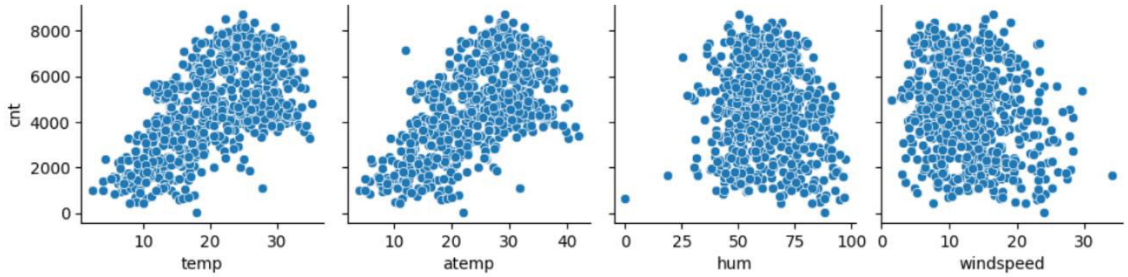
Ans:

- To avoid multicollinearity.
- If we include all levels of a categorical variable as dummy variables, it introduces multicollinearity in the model.
- The categorical variables that have multiple categories, ‘n’ levels and we need ‘n-1’ levels, by get_dummies, when we create dummy variables we convert the categorical data into binary variables.
- For example, if a data point has a value of 0 in all three dummy variable columns created from a categorical variable with three categories, we know it must belong to the category that was dropped using drop_first = True.

- By dropping the first level of each categorical variable when creating dummy variables each level of the categorical variable is compared to a reference level, reducing the risk of multicollinearity.
- In our assignment, after dummy variable encoding:
    - For season variable, 'fall' is the reference category
    - For month variable, 'April' is the reference category
    - For weekdays variable, 'Friday' is the reference category
    - For weather variable, 'clear' is the reference category

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

Ans:



- 'temp' predictor variable has the highest positive correlation with the 'cnt' (target variable) , this suggests that with higher temperatures, more bikes are being rented.

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

Ans:

- To validate linear relationship: We did "plt.scatter(y_test,y_test_pred)" to create a scatter plot of y_test against y_test_pred to visually inspect whether the relationship between the actual and predicted values is linear.
- To validate normality of residuals: We did "sns.distplot(res)" to validate normality. The residuals should ideally follow a bell-shaped curve if they are normally distributed.
- To validate independence of error terms: We need to examine the residuals to ensure there is no systematic pattern or correlation between them.
- To validate homoscedasticity: Plot residuals against predicted values and look for consistent spread of residuals across different predicted values. We did "sns.regplot(x=y_test_pred, y=y_test-y_test_pred)" to validate this assumption for

the test set.

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

Ans: Top three features contributing significantly towards explaining the demand of the shared bikes are 'temp', 'yr' and the 'heavy rain+snow'.

General Subjective Questions

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

Ans:

- Definition: Linear regression is a supervised learning algorithm used to establish relationship between a dependent variable (target) and one or more independent variables (predictor variables).
- The goal is to find a linear equation that best fits the observed data points that would describe the relationship between the independent and the dependent variable, so for this the algorithm calculates the coefficients of the linear equation by minimizing the sum of squared errors between the observed and predicted values. Once the model is trained, it can be used to predict the dependent variable's value based on new input values of the independent variables.
- In simple linear regression, there is only one independent variable. The relationship between the independent variable (X) and the dependent variable (y) is represented by a straight line equation,
  $y = \beta0 + \beta1 * X$
  where,
  $\beta1$ is the slope that represents the change in y for a one-unit change in X and
  $\beta0$ is the intercept that represents the value of y when X is zero.
- In Multiple linear regression, there are multiple independent variables. The relationship between the independent variables (X1, X2, ..., Xn) and the dependent variable (y) is represented by a linear equation,
  $y = \beta0 + \beta1*X1 + \beta2 * X2 + \beta3 * X3 + \ldots\ldots\ldots \beta n * Xn$
- The model can be evaluated using various metrics such as R-squared, mean squared error or adjusted R-squared. These metrics help assess how well the

model fits the data and whether it generalizes well to new data.

- There are few assumptions:
  i) There is linear relationship between x and y: This means that there has to be some linear relationship between the independent variable and the dependent variable.
  ii) Error terms are normally distributed (not x, y) : If the error terms not being normally distributed then the p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable. Visually to prove this assumption it should show a bell shaped curve.
  iii) Error terms are independent of each other: There should be no systematic pattern or correlation between the residuals. This assumption ensures that each data point provides new information and is not influenced by the errors of other data points.
  iv) Constant Variance of Error Terms (Homoscedasticity): The spread of the residuals should be consistent throughout the range of x, to ensure that the model's predictions are equally accurate across different levels of the independent variable.

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

Ans:

- Definition: The Anscombe's quartet is a set of four datasets that have nearly identical statistical properties when analysed using common summary statistics like mean, variance, correlation, and linear regression coefficients. Despite their statistical similarity, the datasets exhibit vastly different patterns when graphically visualized.
- So these four dataset consists of eleven (x,y) points, that share the same descriptive statistics such as mean, variance, standard deviation etc.
- But when we plot then we get different graphs.
  - Below are the different datapoints:

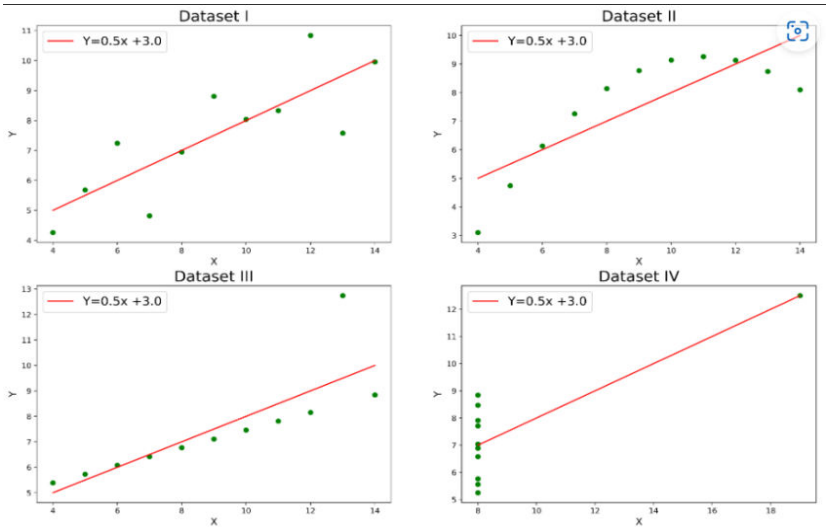
Anscombe's Data

Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

  - If we calculate the statistics, we get:

				Summary Statistics						
N	11	11		11	11		11	11	11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94	3.16	1.94
r	0.82			0.82			0.82		0.82	

- But now when we plot them, we get:



- Form the graphs we can say:
  - Dataset I shows a linear relationship with little variance
  - Dataset II shows a curve and is not linear
  - Dataset III shows somewhat linear but few outliers
  - Dataset IV shows datapoints in a straight line so x is constant
- Thus from this we can understand the importance of data visualization in analysis, instead of solely relying on statistics, we can find relationships, identify if any outliers, etc.

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

Ans:

- It is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. So, this means it will only measure the strength and direction of the linear relationship between two variables.

- It calculates the effect of change in one variable when the other variable changes.

It returns the values between -1 and 1.

- One property of the correlation coefficient is that it remains the same regardless of the units of measurement used for the variables. For example if I have two variables, height in inches and weight in pounds and after calculation R = 0.7, now if I have the same variables, height in centimetres and weight in kilograms, the R will still come out to be 0.7.

- Another property is that the correlation coefficient between the variables is symmetric. For example if one of my variable is X, indicating study hours and another is Y, indicating exam scores, and if R is 0.8 between X and Y, it will still be same between Y and X. Means that the correlation between two variables X and Y is the same regardless of which variable is considered the independent variable and which is considered the dependent variable.

- Pearson correlation coefficient formula:

$$r = \frac{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}{\sqrt{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}}$$

Where,

- x_i and y_i are the individual data points.
- $\bar{x}$ and $\bar{y}$ are the mean value of x and y, respectively.

- From this equation we get the value of r, and we can infer that:
    - If R is close to 1, it indicates a strong positive linear relationship, meaning that as one variable increases, the other variable also tends to increase.
    - If R is close to -1, it indicates a strong negative linear relationship, meaning that as one variable increases, the other variable tends to decrease.
    - If R is close to 0, it indicates a weak or no linear relationship between the variables.

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

Ans:

- By definition, Scaling is a method used to normalize the range of independent variables or features of data; it refers to the process of transforming the values of independent variables to a similar scale.
- We need to perform scaling because when we use variables with varying ranges together in a statistical model like in linear regression, it can cause issues.
- By addressing the varying range issues, we can improve model performance, reduce the impact of outliers, and ensure that the data is on the same scale.
- In machine learning, the model assigns weights to the variables, and the model may prioritize minimizing errors in predicting the output for data points with larger ranges, thus neglecting variables with smaller ranges.
- So if the model assigns large weights to variables with a large range, it may overemphasize their importance in predicting the output, even if they are not actually the most relevant features. This means the model can produce poor results or can perform poorly during learning This can result in poor results and perform poorly during learning, thus decreasing model interpretability.

- There are different scaling techniques:
    A. Standardization (z-score normalization): Standardization scales the data to have a mean of 0 and a standard deviation of 1.
    B. Min-Max scaling: Min-Max scaling scales the data to a fixed range, usually 0 to 1, so min is 0 and at max 1.

Standardization	Min-Max scaling
Mean is 0 and SD is 1	Min value is 0 and Max value is 1
Formula: $z=\{x-u\}/\sigma$ Where: x is the original value μ is the mean of the feature σ is the standard deviation of the feature	Formula: $X\_{new}= (X-X_{min})/(X\_{max}-X_{min})$ Where: X is the original value X_{min} is the minimum value of the feature X_{max} is the maximum value of the feature
Less sensitive to outliers.	Highly sensitive to outliers since min and max are affected by extreme values.
Standardization is often used when the data is assumed to be normally distributed.	Normalization makes no assumption about the underlying data distribution

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

Ans:

- If the value of VIF is infinite, then that means that there is a perfect correlation between the two independent variables or the predictor variables. This will cause a multicollinearity issues, and will affect our analysis.
- VIF assesses how much one independent variable's variance is affected by its relationship with all the other independent variables in the model.
- The VIF is given by:

$$\mathrm{VIF_i} = \frac{1}{1 - \mathrm{R_i^2}}$$

where
VIF_i is the Variance Inflation Factor for the i^{th} predictor variable
R_i^2 is the R^2 value for the i^{th} predictor variable

- Thus if my $R^2 = 1$ then my VIF will be infinite
- So, infinite VIF happens when,
    - If there is perfect collinearity between variables. For example if I have two variables, 'temperature in Celsius' and 'temperature in Fahrenheit', the two variables contain the same information in different units or scale then these two variables are perfectly correlated because one can be converted into the other using a linear equation and if both variables are included in a regression model it will lead to infinite VIF for one of the variables.
    - Another scenario is when, one variable can be perfectly predicted from a combination of the other variables in the model. This is called as linear dependence. For example if I have 'no. of bedrooms' and 'no. of bathrooms' and 'total area', then I can predict the 'total area' from the combination of the other two variables, because the houses that have same combination of bedrooms and bathrooms might have same total area.

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

Ans:

- A Q-Q plot, also known as Quantile-Quantile plot is a graphical tool that is used to

assess whether the data follows a particular probability distribution, such as the normal distribution.

- In a Q-Q plot, we compare the quantiles of the observed data i.e. the residuals from a linear regression model to the quantiles of the theoretical normal distribution.

- The quantiles of the theoretical normal distribution serve as reference points, allowing us to visually assess whether the observed data follows a similar distribution pattern.
- Quantiles of our data are like, for example, if we divide our dataset into four equally sized intervals, the quantiles would be: Q1 , Q2 and Q3, i.e. is the 25th percentile, 50th percentile and 75th percentile respectively.
- And the quantiles of the theoretical distribution are predetermined which are derived from the properties of the normal distribution curve.
- If we are to plot this Q-Q plot then on x-axis is the quantiles of the theoretical distribution and on the y-axis is the quantiles of our actual data.

- What is expected is that, the outcome from a Q-Q plot is to check if the points fall approximately along a straight line, i.e. at a 45 degree angle from the x-axis.
- And if there is deviation from this line, for example, if the points curve upwards, it suggests that the data has heavier tails than the normal distribution.

- Importance in linear regression:
    - To validate the assumption of normality of residuals, in linear regression we assume that the residuals are normally distributed with mean centered at zero and by plotting a Q-Q plot we can visually check if this assumption holds true.
    - Also, we have separate training and test data sets, using Q-Q plot we can confirm that both data sets are from populations with the same distributions.
    - Any deviations from normality in the residuals would suggest that the model is missing important predictors, that the relationship between the predictors and the response is not linear, or that the variability of the response is not constant across different levels of predictors.