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?

Bike sales are increasing in fall, summer & winter season compared to spring season.
Bike sales are better at clear & mist weathersit compared to snow weathersit

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

[answer]
It helps in reducing the extra column created during dummy variable creation . Hence it reduces the correlations created among dummy variables .

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

[answer]
temp & atemp columns

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

[answer]

- The relationship between the dependent and independent variables is linear.
- Low variance inflation factor VIF among (assuming less than 5) independent variables
- Error terms follow a normal distribution, have mean at 0 , have constant variance (homoscedasticity)
- No multicollinearity: The independent variables are not highly correlated with each other.
- There should be no relationship between the errors and the independent variables.
- P value for all independent variables should be less than 0.05.

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

[answer]
Increase in Temperature & year 2019 are positively affecting bike sale count (column ‘cnt’) . Also if the weather is snow then its negatively affecting bike sales count.

In short temp, yr & snow weather are affecting dependent variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

[answer]:

Linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Formula for linear regression

$Y = B0 + B1 * X1 + B2 * X2 + B3 * X3 + \ldots Bp+Xp + E$

Where

B1,B2,.. Bp : Are independent variables co-efficients

X1,X2,...Xp : Are independent variables

B0 : intercept constant

Steps to calculate co-efficients & Intercept are below.

1. Independent values should not have high correlation among them. This we can find out using a pair plot or hist plot. Also we can find correlation by calculating VIF(Variance Inflation factor) Where a variable having higher correlation will have higher VIF value.
2. All Categorical columns have to be converted to numerical columns as Linear Regression can be computed better for continuous values. This can be done by adding dummies columns for all categorical values. Pandas provides standard function to convert Categorical columns to dummy columns
3. Feature Scaling: Another important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead to a model with very weird coefficients that might be difficult to interpret. Scaling can be achieved using standardizing or MinMax scaling.
    - Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.
    - MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.
4. Construct Train set & Test set from the available dataset. Construct the model over Train set.
5. Feature Selection: There are manual & automated approaches. Python provides api to select features based on Recursive Feature Elimination technique(RFE) where one can specify the number of dependent columns we want to keep for constructing our model.

6. Feature selection is a recursive approach. Where we Calculate R-square value & P values & VIF factor for independent variables.

7. Build model(Using OLS method & many other api's python has provided), Drop features that are least helpful in prediction (high p-value), Drop features that are redundant (using correlations, VIF) &Rebuild model and repeat
8. Once the Model is ready for the Train set , verify Error terms.
9. Compare model against Test set & verify R-square values are matching for Train & test dataset & finally construct Linear Regression equation.

2. Explain the Anscombe's quartet in detail.

[answer]:
Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Find the descriptive statistical properties for the all four dataset

- Find mean for x and y for all four datasets.
- Find standard deviations for x and y for all four datasets.
- Find correlations with their corresponding pair of each datasets.
- Find slope and intercept for each datasets.
- Find R-square for each datasets.
    - To find R-square first find residual sum of square error and Total sum of square error
- Create a statistical summary by using all these data and print it.

3. What is Pearson's R?

[answer]:

In statistics, the **Pearson correlation coefficient** (**PCC**)[a] is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive

Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

[answer]:
When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead to a model with very weird coefficients that might be difficult to interpret.
For eg: Support Age, Height & Weight of a person given as Age in Number, Height in Feet & Weight in Pounds. Though values for all 3 variables are numerical in nature, their interpretation is different. To make our calculations independent of these interpretations we need scaling.

So we need to scale features because of two reasons:
1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:
1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

[answer]:
VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.In order to determine VIF, we fit a regression model between the independent variables.

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

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

[Answer]:
Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.
Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.
A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals.