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?

A: The categorical variables from the dataset are 'Season', 'Weathersit', 'Holiday' and 'Year' as these can be clearly spitted into categories

The following can be inferred from the output of OLS regression test about the above categorical variables

Season (We have dropped the variable spring)

1) Rentals of summer is 11.65 percent higher compared to that of spring
2) Rentals of fall is 7.46 percent higher compared to that of spring
3) Rentals of winter is 16.30 percent higher compared to that of spring

This means Rentals are comparatively higher during Winter

Weather (We have dropped the variable clear)

1) Rentals during misty weather is 5.37 lower compared to clear weather
2) Rentals during light rain is 24.09 percent lower compared to clear weather

This means Rentals are comparatively lower during rainy or misty weather. People prefer cycling during clear weather conditions

Holidays

There is an 8.65 percent drop in bike rentals during holidays compared to working days

Year

There is an increase of 23.09 percent in bike rentals in 2019 compared to that of 2018 which indicates an increase in demand over the year

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

A: This helps in preventing multicollinearity and hence helps in creating independent variables

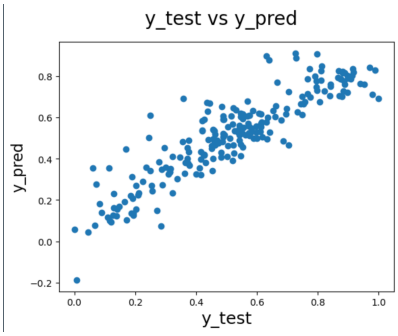
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A: The variables with the highest correlation with the target variable is temp and atemp with temp at 64% and atemp at 65% which is obtained from heatmap and pairplot

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

A: The R2 value of the output is 0.8022 which indicates a strong fitment of variables.

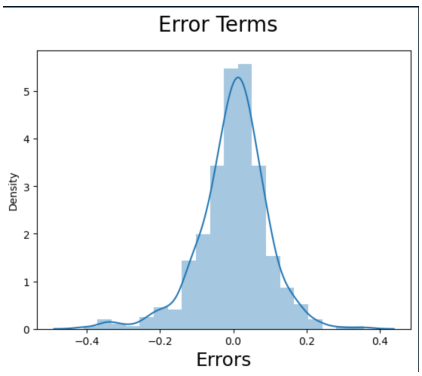
The plot between y_test and y_pred is also shows a strong linear relationship between the predicted and actual values



The VIF values of all the variables except the target variable are under 5 as given below

	Features	VIF
0	const	44.87
1	yr	1.03
2	holiday	1.01
3	temp	3.50
4	hum	1.87
5	windspeed	1.19
6	summer	2.54
7	fall	4.78
8	winter	1.87
9	mist	1.56
10	light rain	1.24

It is also seen from the notebook that the errors are normally distributed



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

A: Temperature with a coefficient of .4961

Year with a coefficient of .2307 and

Light Rain with a coefficient of -.2409

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A: Simple linear regression is a statistical method for establishing the relationship between two variables using a straight line. The line is drawn by finding the slope and intercept, which define the line and minimize regression errors.

The simplest form of simple linear regression has only one x variable and one y variable. The x variable is the independent variable because it is independent of what you try to predict the dependent variable. The y variable is the dependent variable because it depends on what you try to predict.

$h(x) = y = \beta_0 + \beta_1 x + e$ is the formula used for simple linear regression.

y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).

β_0 is the intercept, the predicted value of y when the x is 0.

β_1 is the regression coefficient – how much we expect y to change as x increases.

x is the independent variable (the variable we expect is influencing y).

e is the error of the estimate, or how much variation there is in our regression coefficient estimate.

Simple linear regression establishes a line that fits our data, but it does not guarantee that the line is good enough. For example, if our data points have an upward trend and are very far apart, then simple linear regression will give you a downward-sloping line, which will not match our data.

When predicting a complex process's outcome, it's best to use multiple linear regression instead of simple linear regression. But it is not necessary to use complex algorithms for simple problems.

A simple linear regression can accurately capture the relationship between two variables in simple relationships. But when dealing with more complex interactions that require more thought, we need to switch from simple to multiple regression.

A multiple regression model uses more than one independent variable. It does not suffer from the same limitations as the simple regression equation, and it is thus able to fit curved and non-linear relationships.

The formula for multiple regression is

$h(x_1,x_2,...,x_k)=\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k$

where, $x_1,x_2,...,x_k$ are the independent variables.

β_0 is the intercept.

$\beta_1,\beta_2,...,\beta_k$ are the coefficients, representing the influence of each respective independent variable on the predicted output.

The following are the assumptions taken in case of simple linear regression

Linearity: The relationship between x and y should be linear. It means that, as one value increases, the other increases correspondingly. We can use a scatterplot to show this linearity.

Independent of Errors: It is essential to check if our data is independent of errors. If there is a relationship between the residuals and the variable, this could cause problems with our model. To check the independence of errors, we can examine a scatterplot of “residuals versus fits”; it should not look like there is a relationship.

Normal Distribution: It is also essential to check if our data is normally distributed. To do this, we have to examine a histogram of the residuals. It should be approximately normally distributed. The histogram should also show that most of your observations are close to 0 or 1 (the max/min values). It will help us to make sure that our model is accurate and reliable.

Variance Equality: Finally, it is essential to check if our data have equal variances. To do this, we can examine a scatterplot and look for any outliers or points that seem far from each other in conflict. If there are outliers or points with high variance compared to others.

It is not easy to get the best fit line in real life cases so we need to calculate errors that affects it. These errors need to be calculated to mitigate them. The difference between the predicted value $\hat{Y}$ and the true value Y and it is called cost function or the loss function.

In Linear Regression, the Mean Squared Error (MSE) cost function is employed, which calculates the average of the squared errors between the predicted values $\hat{y}_i$ and the actual values y_i. The purpose is to determine the optimal values for the intercept θ_1 and the coefficient of the input feature θ_2 providing the best-fit line for the given data points. The linear equation expressing this relationship is $\hat{y}_i = \theta_1 + \theta_2 x_i$.

The MSE function can be calculated as follows:

$$\text{Cost function}(J) = \frac{1}{n}\sum_{n}^{i} (\widehat{y_i} - y_i)^2$$

Utilizing the MSE function, the iterative process of gradient descent is applied to update the values of θ_1 and θ_2. This ensures that the MSE value converges to the global minima, signifying the most accurate fit of the linear regression line to the dataset.

This process involves continuously adjusting the parameters θ_1 and θ_2 based on the gradients calculated from the MSE. The result is a linear regression line that minimizes the overall squared differences between the predicted and actual values, providing an optimal representation of the underlying relationship in the data.

After calculating the loss function, the next step is to optimize model to mitigate this error. This is done through gradient descent.

Gradient descent is an optimization technique used to train a linear regression model by minimizing the prediction error. It works by starting with random model parameters and repeatedly adjusting them to reduce the difference between predicted and actual values.

Algorithm for gradient descent can be represented as follows

- Start with random values for slope and intercept.
- Calculate the error between predicted and actual values.
- Find how much each parameter contributes to the error (gradient).
- Update the parameters in the direction that reduces the error.
- Repeat until the error is as small as possible

This helps the model find the best-fit line for the data.

Evaluation Metrics for Linear Regression

A variety of evaluation measures can be used to determine the strength of any linear regression model. These assessment metrics often give an indication of how well the model is producing the observed outputs.

The most common measurements are:

1. Mean Square Error (MSE)

Mean Squared Error (MSE) is an evaluation metric that calculates the average of the squared differences between the actual and predicted values for all the data points. The difference is squared to ensure that negative and positive differences don't cancel each other out.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Here,

n is the number of data points.

y_i is the actual or observed value for the i^{th} data point.

($\hat{y}_i$ is the predicted value for the i^{th} data point.

MSE is a way to quantify the accuracy of a model's predictions. MSE is sensitive to outliers as large errors contribute significantly to the overall score.

2. Mean Absolute Error (MAE)

Mean Absolute Error is an evaluation metric used to calculate the accuracy of a regression model. MAE measures the average absolute difference between the predicted values and actual values.

Mathematically MAE is expressed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | Y_i - \hat{Y}_i |$$

Here,

n is the number of observations

$Y_i$represents the actual values.

$\hat{Y}_i$ represents the predicted values

Lower MAE value indicates better model performance. It is not sensitive to the outliers as we consider absolute differences.

3. Root Mean Squared Error (RMSE)

The square root of the residuals' variance is the Root Mean Squared Error. It describes how well the observed data points match the expected values or the model's absolute fit to the data.

In mathematical notation, it can be expressed as:

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\frac{\sum_{i=2}^{n}(y_i^{actual} - y_i^{predicted})^2}{n}}$$

Rather than dividing the entire number of data points in the model by the number of degrees of freedom, one must divide the sum of the squared residuals to obtain an unbiased estimate. Then, this figure is referred to as the Residual Standard Error (RSE).

In mathematical notation, it can be expressed as:

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\frac{\sum_{i=2}^{n}(y_i^{actual} - y_i^{predicted})^2}{(n-2)}}$$

RSME is not as good of a metric as R-squared. Root Mean Squared Error can fluctuate when the units of the variables vary since its value is dependent on the variables' units (it is not a normalized measure).

4. Coefficient of Determination (R-squared)

R-Squared is a statistic that indicates how much variation the developed model can explain or capture. It is always in the range of 0 to 1. In general, the better the model matches the data, the greater the R-squared number.

In mathematical notation, it can be expressed as:

$$R^2 = 1 - (\frac{RSS}{TSS})$$

Residual sum of Squares(RSS): The sum of squares of the residual for each data point in the plot or data is known as the residual sum of squares or RSS. It is a measurement of the difference between the output that was observed and what was anticipated.

$$RSS = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Total Sum of Squares (TSS): The sum of the data points' errors from the answer variable's mean is known as the total sum of squares or TSS.

$$TSS = \sum_{i=1}^{n} (y - \bar{y_i})^2.$$

R squared metric is a measure of the proportion of variance in the dependent variable that is explained the independent variables in the model.

5. Adjusted R-Squared Error

Adjusted R^2 measures the proportion of variance in the dependent variable that is explained by independent variables in a regression model. Adjusted R^2 the number of predictors in the model and penalizes the model for including irrelevant predictors that don't contribute significantly to explain the variance in the dependent variables.

Mathematically, adjusted R^2 is expressed as:

$$AdjustedR^2 = 1 - (\frac{(1 - R^2).(n - 1)}{n - k - 1})$$

Here,

n is the number of observations

k is the number of predictors in the model

R^2 is coeeficient of determination

Adjusted R^2 helps to prevent overfitting. It penalizes the model with additional predictors that do not contribute significantly to explain the variance in the dependent variable.

While evaluation metrics help us measure the performance of a model, regularization helps in improving that performance by addressing overfitting and enhancing generalization.

Lasso Regression is a technique used for regularizing a linear regression model, it adds a penalty term to the linear regression objective function to prevent overfitting.

The objective function after applying lasso regression is:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y_i} - y_i)^2 + \lambda \sum_{j=1}^{n} | \theta_j |$$

The first term is the least squares loss, representing the squared difference between predicted and actual values.

The second term is the L1 regularization term, it penalizes the sum of absolute values of the regression coefficient θj.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help us in identifying the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

We can define these four plots as follows:

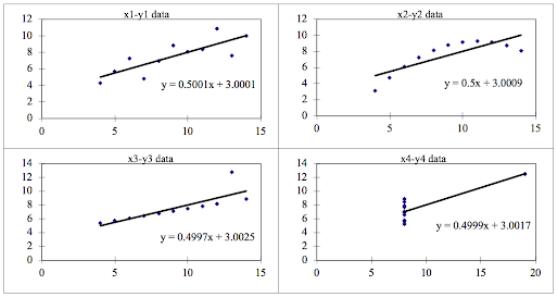
(The following data set is pulled from an online repository)

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

The statistical information for these four data sets are approximately similar. We can compute them as follows:

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
			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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



Anscombe's Quartet Four Datasets

- Data Set 1: Fits the linear regression model well.
- Data Set 2: Cannot fit the linear regression model because the data is non-linear.
- Data Set 3: Shows the outliers involved in the data set, which cannot be handled by the linear regression model
- Data Set 4: Shows the outliers involved in the data set, which also cannot be handled by the linear regression model

The above set of data and its visual representation helps us in clearly understanding how important it is to represent data visually and proves that it is important to view data visually before we get into regression

3. What is Pearson's R?

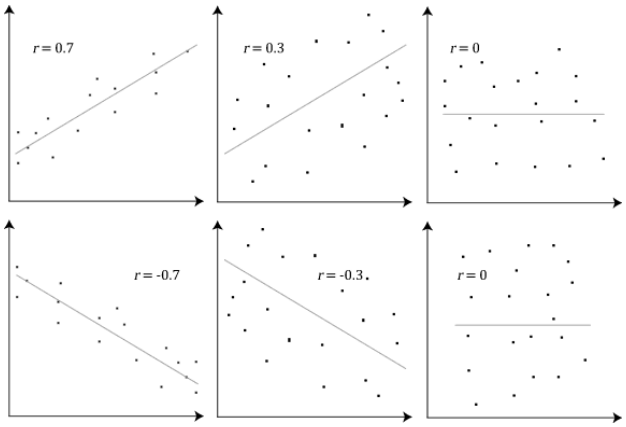
The Pearson product-moment correlation coefficient (or Pearson correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by r. Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient 'r' indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

It can take a range of values from -1 to +1

A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit. The following visual representation contains a r represented against a group of data with positive correlation, negative correlation and no correlation

Visual representation:



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

Scaling is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor.

Scaling law, a law that explains how many natural phenomena exhibit scale invariance.

Scaling performed because it is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. Most of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range.

The difference between normalized scaling and standardized scaling

The values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

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

Variance Inflation Factor (VIF) becomes infinite when a predictor variable is perfectly linearly correlated with one or more other predictor variables in the model.

Why does this happen?

VIF for a predictor X_i is defined as:

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

where R_i^2 s the R-squared value obtained by regressing X_i on all the other predictors.

When does VIF $\rightarrow \infty$

If the variable X_i can be exactly predicted from other predictors, then:

$$R_i^2 = 1$$

Substituting this:

$$VIF_i = \frac{1}{1 - 1} = \frac{1}{0} = \infty$$

This means there is perfect multicollinearity and hence VIF $\rightarrow \infty$

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

Q-Q plot, short for quantile-quantile plot, is a scatterplot that compares the quantiles of two distributions. One distribution is usually the observed data, and the other is a theoretical or reference distribution, such as the normal distribution. The idea is to see how well the data fit the expected distribution by checking if the points lie on or near a straight line.

To create a Q-Q plot, we need to sort the data from smallest to largest and assign them ranks. Then, we need to calculate the expected quantiles of the reference distribution for each rank. For example, if we use the

normal distribution, we can use the inverse cumulative distribution function (CDF) to find the expected quantiles. Finally, we need to plot the observed data on the y-axis and the expected quantiles on the x-axis.

To interpret a Q-Q plot, we need to look at the shape and pattern of the points. If the points lie on or close to a 45-degree line, it means that the data follow the reference distribution closely. If the points deviate from the line, it means that there are some differences between the data and the reference distribution. For example, if the points are curved, it means that the data are skewed or have heavy tails. If the points are scattered or have gaps, it means that the data have outliers or are multimodal.

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, we 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. We can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, we need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.