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)

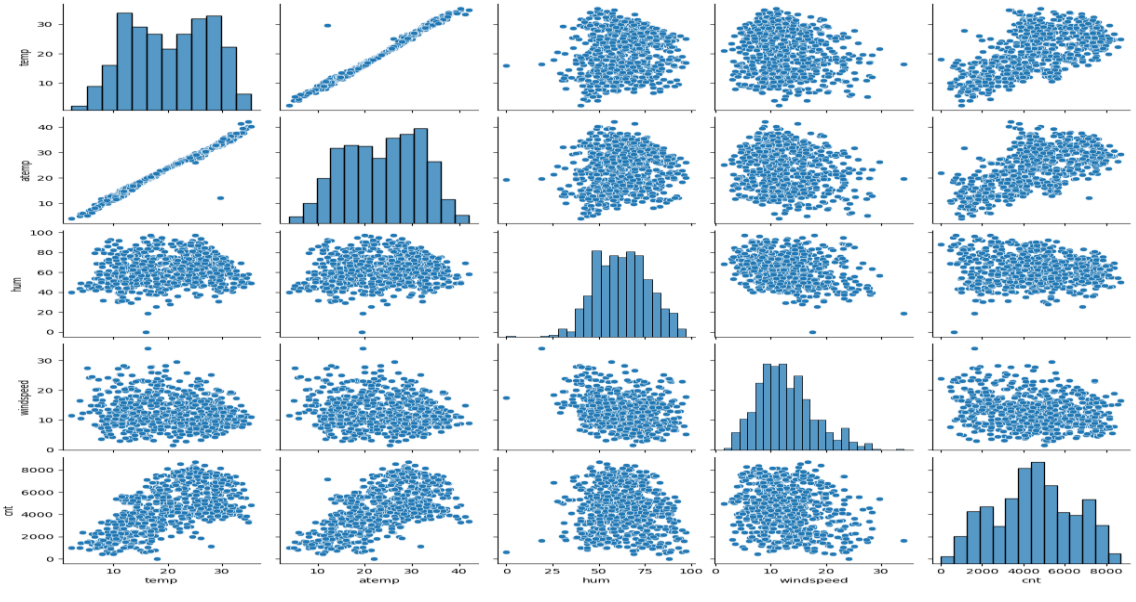
   - Year shows that year by year number of users have been increased. So, post pandemic user count might increase.
   - More number of users on holiday, weekday further shows that especially on Saturday's the user count is high as compared to Sunday.
   - Weathersit shows clear weather has high number of users.
   - More number of users in Fall season, this information is supplemented by month wise user count where a greater number of users in the month of September and October

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

   - N category information can be represented by N-1 categories. i.e every N-1 category is represented by 1, while the last category is represented by 1 the other N-1 category takes 0 value. Instead of explicitly representing the last category, it can be dropped, and when all the N-1 category takes 0 value, it should be interpreted as dropped category is being represented.

   - This approach is to reduce the number of variables while building a linear model and there by the calculation of cost function becomes easier.

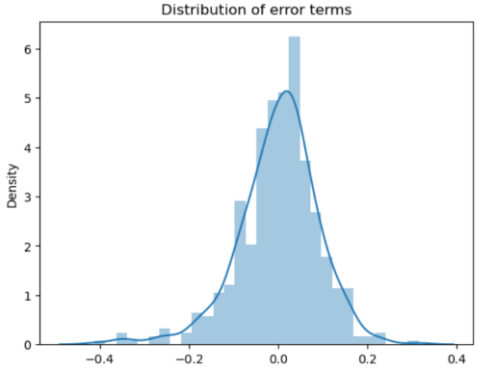
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



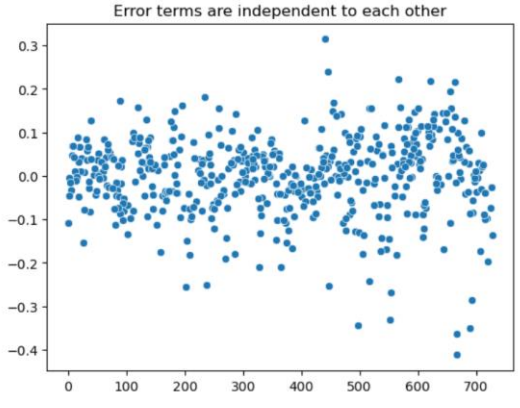
   - Temp and atemp variable has the highest correlation.

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

- The error terms i.e. difference between y train set and y predicted are normally distributed.
- The error terms are distributed normally with mean equal to ZERO.



- Ther error terms are independent to each other.



- The error terms have constant variance.

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 coefficient of 0.5499
    b. Year with coefficient of 0.2331
    c. Light_Snow_Rain_Thunder with coefficient of -0.2871 – i.e with every unit increase of X, y value decreased by 0.2871.

General Subjective Questions

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

Linear regression algorithm shows a linear relationship between input variables (independent/predictive variables) vs output variable (dependent variable). Here when the predicted output variable is continuous and numerical in nature, it's called regression. This algorithm is classified as supervised learning method.

To predict the output, a straight line is fitted, and the equation of this line is given by y=mx+c

Where:

m is gradient (slope) – this tells the how much change in independent variable affects the change in Y variable.

c is the intercept on Y

Further, the line is being made to fit in such way the predicted output has minimal residual error. So, this is calculated by subtracting the actual y point with y predicted points. The difference is squared and summed. This method is called Ordinary Least square method.

e = y – y pred

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

RSS = residual sum of squares

y_i = i^th value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

This best fit is obtained by minimizing the cost function and here cost function is RSS. The cost function is minimized with two approaches:

1. Differentiation
2. Gradient decent – This is done either by minimizing the cost function.

There are certain limitations with Residual Sum of Squares. They are affected by units used by input and output variable. For ex: If **X** is in **lakhs** and output **y** is in **crores**, then RSS value will be different as opposed when both input and output are in **lakhs**. To handle this, $\mathbf{R^2}$ value is used which is the ratio of RSS and TSS

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

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

Further, there are certain assumptions for the error terms i.e. y – y pred in linear regression:

- Error terms are independent to each other.
- Error terms have constant variance.
- Error terms are normally distributed with mean zero.
- Also, X and Y should have linear relationship.

Classification of Linear regression:
Linear regression is classified as simple linear regression and multiple linear regression.

Simple linear regression: It has single input variable i.e. independent variable vs single output/dependent variable. This is given by:-

$y_i=\beta_0+\beta_1 x$

β_0 - This is the intercept.

$\beta_1 x$ - X is independent variable.

Multiple linear regression: As the name suggests, it has multiple independent variables as input vs single output/dependent variable. This is given by:-

$y_i=\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+...+\beta_k x_{i,k}+\epsilon_i.$

β_0 – This is the intercept.

$\beta 1 x_{i,1}$ – Xi is independent variable

$\beta 2 x_{i,2}$ - Xi is independent variable, so on

This needs to be interpreted as when X1 increases, expected value of Y increases by $\beta 1 x 1$ provided all the other predictors are held constant.

With Multiple linear regression, some new aspects need to be considered:

- Overfitting: As we add a greater number of variables, the model becomes too complex and ends up memorizing the training data and fails to generalize. The model is set to overfit when the training accuracy is too high.

- Multicollinearity: Associations between independent variables. Having multiple independent variables which are having high correlation makes the model redundant. So, it is advised to check the VIF score among all the independent variables and score should be less than 5.

- Feature selection: Selecting optimal set of features from the pool variables is an important task to build a model.

Further, When the independent variables count is too high, manual approach of selecting each variable is time consuming and complexity increases. So, we have few methods,

Example: Recursive Feature selection

Model Selection:

Finally, when we have multiple models over a given dataset, we need to keep the model simple and explaining the highest variance. This can be done by penalizing the models having large numbers of predictor variables. This is done by following algorithms:

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

$$AIC = n \times log(\frac{RSS}{n}) + 2p$$

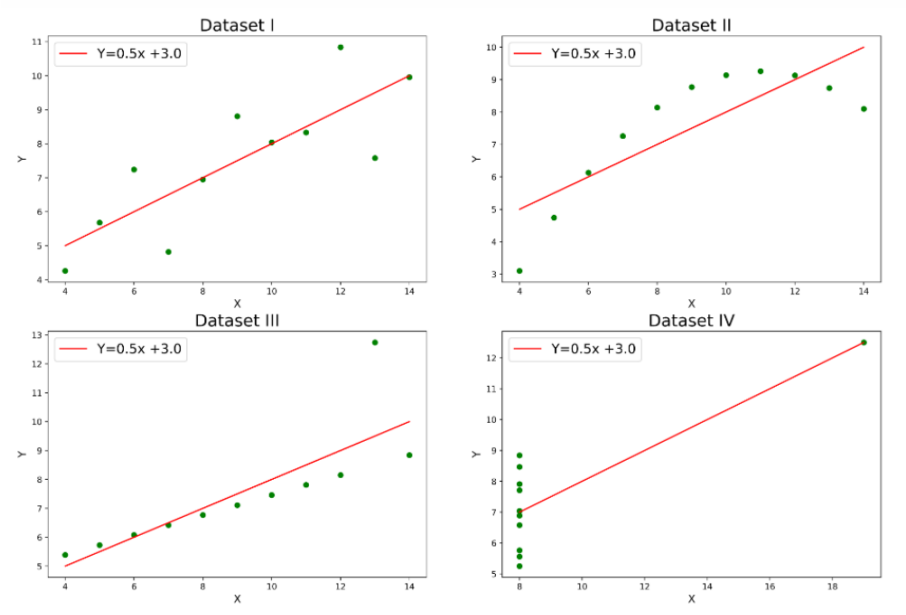
Here by concluding the briefing of linear regression algorithm.

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

Anscombe's quartet consists of 4 different data sets all having identical ***mean*** and ***standard deviation***. This gives us an intuition of the data being similar but when the same data is visualized over a scatter plot the results are different.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Above 4 data show similar results but when visualized over a scatter plot, it gives different results.



Anscombe's quartet Plot

Ref: geeksforgeeks

This reveals the limitations of summary statistics, emphasizing the need for visual exploration to detect nuances, outliers, and diverse relationships in datasets.

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

The Pearson coefficient is a measure of the strength of the association between two continuous variables. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

The Pearson coefficient shows correlation, not causation.

Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

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 normalizing or standardizing the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform scaling.

Scaling helps in faster computation – faster convergence for gradient decent in large data sets. Scaling doesn't impact p-value, r2 or f-statistic, only co-efficient will be impacted.

Scaling methods:

a. Normalized scaling - Normalization means the scaling down of the data set such that the normalized data falls between 0 and 1. This technique compares the corresponding normalized values from two or more different data sets discarding the various effects in the data sets on the scale, i.e., a data set with large values can be easily compared with a smaller values dataset.

$$\text{X normalized} = \frac{(\text{X} - \text{X minimum})}{(\text{X maximum} - \text{X minimum})}$$

b. Standardized scaling – Here the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

   This is obtained by dividing the value by the standard deviation after the mean has been subtracted.

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

VIF – Variation inflation factor indicates how much collinearity between independent variables exists. If independent variable can be described by other independent variables, then it has perfect correlation and has on R squared value of 1. So, as per the formula VIF = $1 /(1 - R2)$, when R^2 is 1 will give 1/0 which is infinity. So, we can conclude that if there is high correlation between independent variables, then VIF can be infinite.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Its tool for comparing the shapes of different distributions. A scatter plot generated by two sets of quantiles against each other is known as Q-Q plot.

As both sets of quantiles come from the same distribution, the points should form a line.

The q-q plot is used to answer the following questions:

Do two data sets come from populations with a common distribution?
Do two data sets have common location and scale?
Do two data sets have similar distributional shapes?
Do two data sets have similar tail behaviour?