Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Here is the analysis of categorical variable on the dependent variable `cnt`

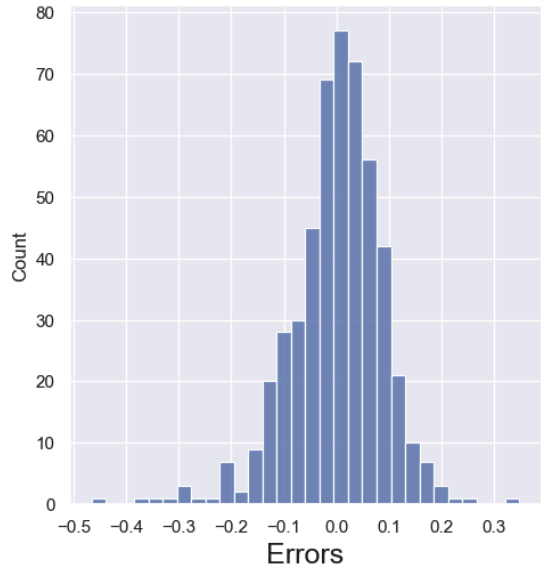
- **Year** - There is a growing number of users from 2019 to 2019
- **Season** - `spring` season has least number of rental users and `fall` has highest number of users whereas `summer` and `winter` has relatively good usage
- **Months** - To validate the observations made in season categories, `Jan` and `Feb` have least rental users and gradually increase over the months and peaks at `August`.
- **Weekday** - Weekday doesn't influence the rental behaviour as we can see similar usage across all the days in a week
- **Holiday** - We have very few users during holidays. ~98% of the usage is during non-holidays
- **Working Days** - ~70% of the usage is during working days.
- **Weather** - Clear days have higher usage of bikes whereas light rainy days have very less usage. We don't have information about thunderstorm days in the dataset.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)
It is important to use drop_first=True because it reduces the number of extra columns created. Since the first column can be explained by having `0` in all other columns the first column becomes unnecessary. Also having that might introduce higher correlations amongst dummy variables created. Having multicollinearity can lead to unstable and unreliable coefficient estimates.

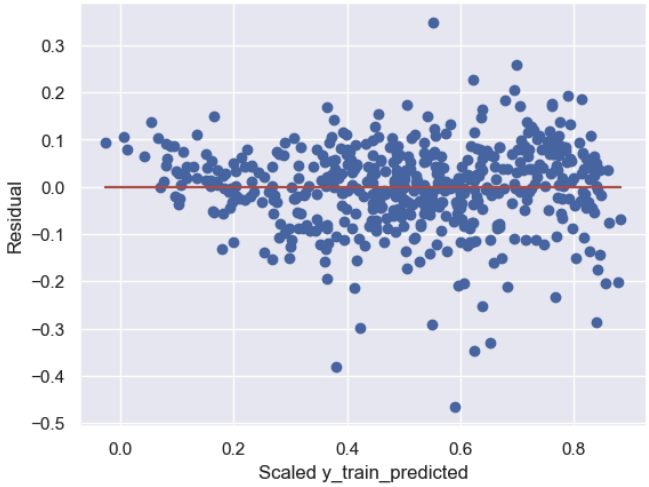
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)
`temp` has the higher correlation with the target variable `cnt`

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. Validate Normal distribution of errors - Plotted distplot of error (i.e predicted - actual) vs count and validated that the graph is normally distributed with peak at 0.



2**. homoscedasticity -** To validate homoscedasticity i.e the variance is independent of the error value, plotted a scatter plot of residual vs predicted value. There is no clear pattern in the scatter plot and also the variance is constant



3. Multicollinearity - The VIF value should not be significantly high

4. Adjusted R-squared value - The adjusted R-squared value should not be very low compared to R-squared value to ensure we have considered only the required variables. Also higher adjusted R-Squared value also indicates higher linearity of dependent vs independent variables.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

1. temp which has 0.38 as coefficient

2. light_rain weather which as negative correlation of -0.29
3. yr has 0.23 as coefficient

General Subjective Questions

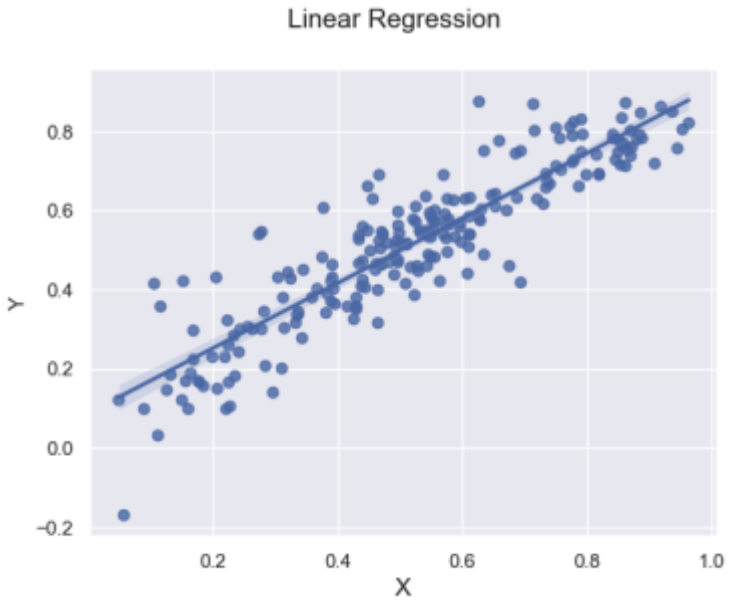
Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression algorithm is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and independent features. When there is only one independent feature, then we call as **Simple Linear Regression** and if there are more features, then we call it as **Multiple Linear Regression**

The Simple Linear Regression takes the equation of y = mx + c where
c = intercept i.e the value of y when x = 0,
m = slope which shows the degree of relationship. Higher the slope, higher change in `y` value with unit change in x.

The Simple Linear Regression takes the equation of $y = m_0x_0 + m_1x_1 + m_2x_2 + m_3x_3 + \ldots. + m_nx_n + c$ where
$m_0, m_1, \ldots m_n$ are the coefficients of each independent variable which tells the degree of weightage of the corresponding independent variable of the dependent variable when all other independent variables are kept constant.

The aim of the algorithm is to find a **best fit line** (a linear line in the below diagram) that represents the linear relationship between the dependent variable and independent features on observed data that has minimal overall error between the actual and the predicted values.



To fit the best fit line, we utilise the cost function to compute the best values for intercept and coefficients. The cost function we use here is **MSE (Mean Squared Error)** which is

$$\frac{1}{n}\sum_{n}^{i}(\widehat{y}_{i} - y_{i})^{2}$$

By using the MSE function, we apply the gradient **descent algorithm** to gradually update the values of m_0 m_1 m_2 ..... m_n and c to identify the equation that has minimal MSE value.

Linear regression Algorithm is based on few assumptions

1. **Linearity**: The independent and dependent variables have a linear relationship with one another.

2**. Error terms are independent of each other**: The error terms should not be dependent on one another.

3**. homoscedasticity -** The variance should not either change or follow a pattern as the error value changes. This indicates that the amount of independent variables has no impact on the variance of the errors

4. **Error terms are normally distributed** - The error terms should be normally distributed with mean equal to 0

5. **Multicollinearity** - The independent features should not be correlated. Otherwise it will be difficult to determine the individual effect of each variable on the dependent variable.

6. **Overfitting** - the model should not be overfit i.e the case where having large independent variables makes the model learn and memorise each and every training dataset which leads to higher accuracy on train dataset but poor accuracy on test dataset. The variables should be carefully selected to avoid making the model complex and overfit.

7. **Feature Selection** - the algorithm runs well when the redundant variables are not included in the model training resulting in avoiding the multicollinearity or overfitting problems

Once we find the best fit line, we evaluate the confidence in the model by checking summary statistics like R-squared value, AIC, Probability of F-Statistic etc and also validate all the assumptions listed above.
Once the model is successfully validated, we evaluate the model against the test data to make sure the model works well on the unseen test data as well.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is four dataset having similar summary statistics but different representation when we scatter plots on a graph. The quartet is 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.

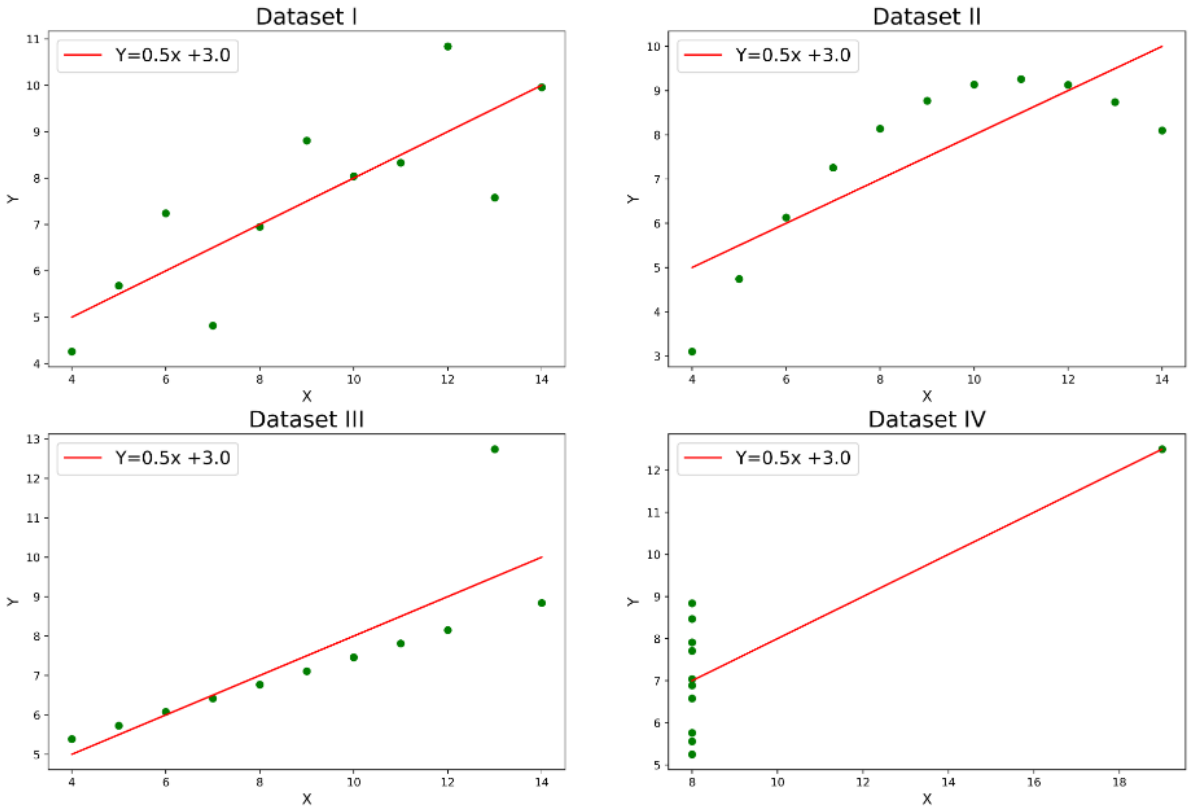
Following is the Anscombe's quartet dataset

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

And here is the summary statistics for all four dataset

- Mean_x = 9.0
- Variance_x = 11.0
- Mean_y = 7.5
- Variance_y = 4.12
- Correlation = 0.816
- Slope = 0.5
- Intercept = 3.000

This could make us think that each dataset is identical or similar in nature. But if we try to create scatter plots for the above dataset, it shows the difference in the dataset clearly even though they have the same LR fit line



- Clearly Dataset 3 demonstrates a higher linear relationship between x and y.
- But Dataset 1 tells some sort of linear relationship but not as strong as dataset 3
- Dataset 2 doesn't seem to have linear relationship
- Dataset 4 shows that one point is enough to produce higher correlation coefficient to assume that the data has linear relationships.

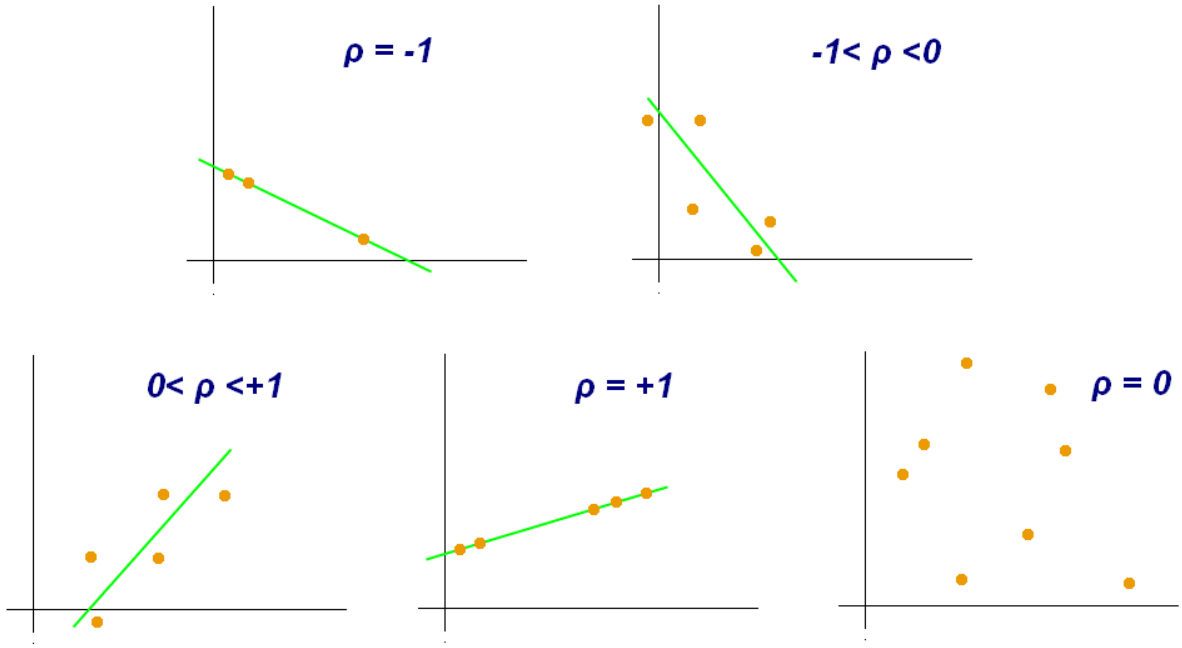
Hence we should solely not rely on the summary statistics instead we also need to look at the visualisations to understand the relationship between X and y.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R also known as Pearson's coefficient is the descriptive statistics that tells the strength of relationships between two variables and also the direction (i.e do they move in the same direction or opposite direction).
The value ranges between -1 to 1 where

- -1 tells the variables are strongly related but in opposite direction (one increases and other decreases)
- +1 tells the variables are strongly related in same direction (both increases or decreases together)
- 0 indicates not related to each other

Following graphs show different scatter plots for different pearson's r values.



Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

There are two ways of scaling

- Normalised scaling (Min-max scaling) - where all the independent variables are normalised between the range 0 and 1 i.e [0,1]
- Standardized scaling - all independent variables are scaled around mean 0 and have unit variance

Why scaling is performed -

- Gradient descent based algorithm - If an algorithm uses gradient descent, then the difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model
- Distance based algorithms - Distance-based algorithms like KNN, K-means, and SVM are most affected by the range of features. This is because behind the scenes they are using distances between data points to determine their similarity and hence perform the task at hand. Therefore, we scale our data before employing a distance-based algorithm so that all the features contribute equally to the result.

Difference between Normalised and Standardised scaling:

Normalisation	**Standardisation**
This method scales the model using minimum and maximum values.	This method scales the model using mean and standard deviation
When features are on various scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

VIF tells the degree of multicollinearity among the independent variables. Higher the VIF value, greater is the multicollinearity. VIF value 1 tells the independent variables are not related to each other.

VIF = $\frac{1}{1-R^2}$

In the case of perfect multicollinearity between independent variables, the R-squared value will be 1 and hence VIF will become infinity. If we get infinity as VIF value, then we need to remove one of the variables that are perfectly correlated.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The quantile-quantile(Q-Q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not.
It is used to check if two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with the same distributions.