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

- From 2018 to 2019 there is increase in bikes hiring and which show there is improvement in industry
- There will be equal hiring of bikes on working and non working day.
- Most of the booking done, where there are temperature between range of 20 to 30 c. Which shows people avoid to go out in cold and hot season
- Most booking done in May, June, July, August, September, and October
- 2019 is higher booking year, which shows down the line there is positive impact on the business for BoomBike.
- If whether is clear , hiring rate for bikes are high.

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

- If we do not use drop_first = True, then the dummy variables will be created, and these predictors are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.
- It helps in reducing the extra column created during dummy variable creation.
  And dropping that column is necessary as it will them impact overall correlation among variables.
- By creating n − 1 dummy variables, you avoid perfect mul>collinearity, as the information about the ommited category is implicitly captured.
- In Python, when using libraries like pandas, we can set `drop_first=True` while creating dummy variables to automatically drop one of the dummy variables to adhere to this n-1 rule

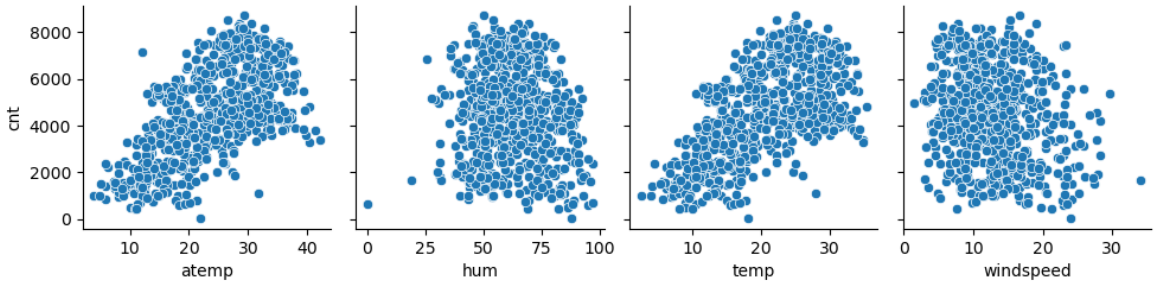
---

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

The variable 'temp' exhibits the strongest correlation with the target variable, as depicted in the graph below. Given that 'atemp' and 'temp' are redundant variables, only one of them is selected for the best fit line.



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

- For model to be fit in linear Regression, we have to look for following
- It must have lowest p-value for all the categorial variable, which will ensure that model is prone to be perfect

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.849
Model:	OLS	Adj. R-squared:	0.844
Method:	Least Squares	F-statistic:	185.1
Date:	Mon, 28 Oct 2024	Prob (F-statistic):	7.37e-192
Time:	23:00:49	Log-Likelihood:	-4114.1
No. Observations:	511	AIC:	8260.
Df Residuals:	495	BIC:	8328.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4491.3033	34.117	131.646	0.000	4424.272	4558.334
yr	1002.7700	34.731	28.873	0.000	934.533	1071.007
holiday	-136.0693	34.751	-3.916	0.000	-204.346	-67.793
temp	985.0134	67.981	14.489	0.000	851.446	1118.581
hum	-202.6100	47.198	-4.293	0.000	-295.342	-109.878
windspeed	-270.7115	37.538	-7.212	0.000	-344.465	-196.958
Spring	-205.3954	77.805	-2.640	0.009	-358.265	-52.526
Summer	181.1941	55.884	3.242	0.001	71.394	290.994
Winter	377.8064	66.239	5.704	0.000	247.663	507.950
January	-86.2534	42.739	-2.018	0.044	-170.226	-2.281
July	-122.6281	41.654	-2.944	0.003	-204.469	-40.788
November	-61.8901	41.575	-1.489	0.137	-143.575	19.795
September	185.6439	39.002	4.759	0.000	108.838	262.450
Monday	-138.8696	34.447	-4.031	0.000	-206.550	-71.190
Good	1046.0370	109.463	9.556	0.000	830.968	1261.106
Moderate	793.4624	101.812	7.793	0.000	593.425	993.500

Omnibus:	64.888	Durbin-Watson:	2.061
Prob(Omnibus):	0.000	Jarque-Bera (JB):	164.270
Skew:	-0.648	Prob(JB):	2.13e-36
Kurtosis:	5.457	Cond. No.	6.81

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

As you can see in above diagram, p value for November is high. Hence, we have to build model again with elimination of November from Dataset.

- We have to do check on F-statistic value as well to fit in with regression
- VIF Variance Inflation Factors : Value for VIF is less than 5, and to achieve that we have to

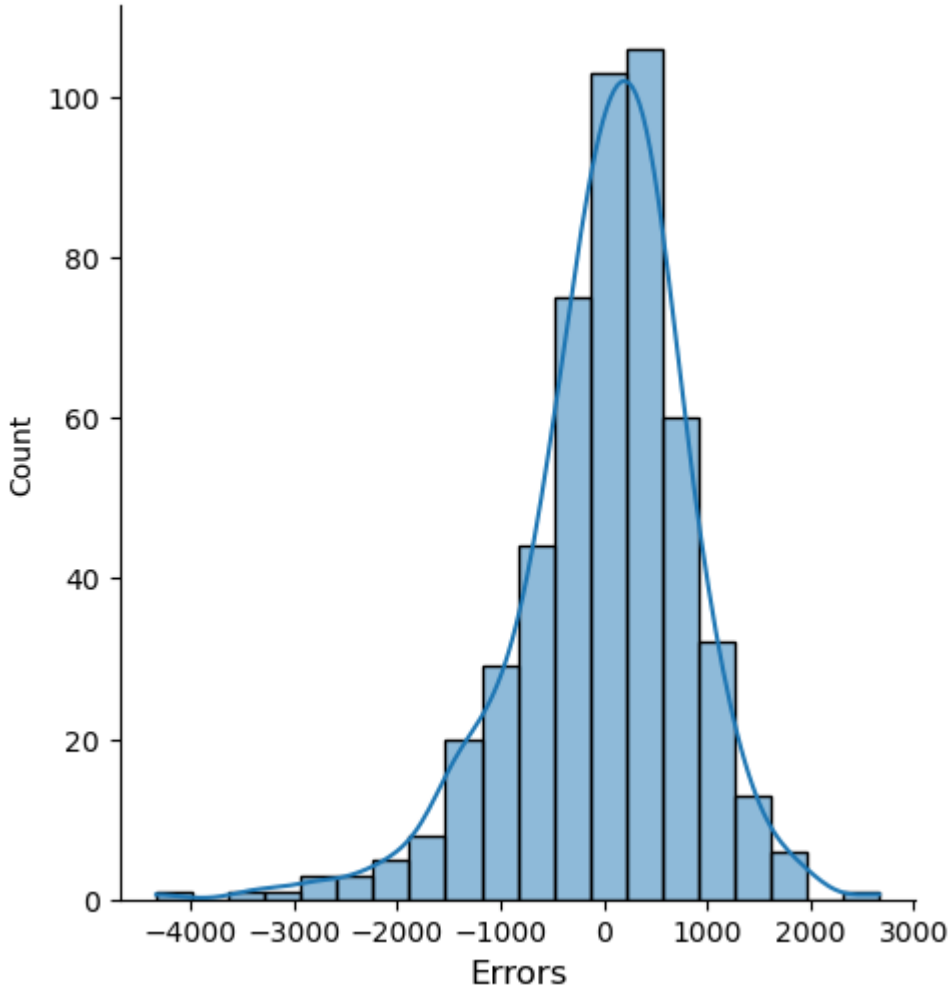
eliminate those variable which are having higher VIF (more than 5)

-

	Features	VIF
0	yr	1.04
1	holiday	1.04
2	temp	3.97
3	hum	1.91
4	windspeed	1.21
5	Spring	5.20
6	Summer	2.68
7	Winter	3.77
8	January	1.57
9	July	1.49
10	November	1.49
11	September	1.31
12	Monday	1.02
13	Good	10.29
14	Moderate	8.91

As you can see in above diagram, we have to eliminate **Good** feature.

- We must avoid overfitting, as sometime it is due to invalid entries in Dataset.
- Examine residuals for autocorrelation



➢

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

- After building model , below is final equation :
- cnt = 4491.30 + 992.43 x yr + 1229.32 x temp -442.38 x hum - 350.59 x windspeed + 303.88 x Summer + 510.12 x Winter -123.63 x July + 211.14 x September -109.19 x Monday
- Three main features that are impacting.
    - **Temperature**
    - **Winter season**
    - **Year**

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 is a statistical method used for creating model and which is nothing but relationship between a dependent variable and one or more independent variables. By use of linear regression we are actually predicting value of dependent variable and impact of independent variable on it.

We have to find best fitting line that minimizes the sum of the squared differences between the observed and predicted values of the dependent variable.

Linear regression is used for prediction and forecasting, and is often used in machine learning

Dependent variable: Variable that you want to predict

Independent variable: Variable that you use to predict the dependent variable

Linear equation: Equation that is used to best fir your prediction

y: the dependent variable

x: Independent variable

b: The Y intercept

a: the slop of the line

e: represents error term

From above parameter equation:

y= b + ax + e

If we have more than one independent variable

y= b + a1x1+a2x2….+anxn + e

While building model we have to use MSE,

The goal is to find the values of (a1, a2, a3, ...an) that minimize the sum of the squared differences between the observed and predicted values. This is often expressed as the sum of squared errors (SSE) or mean squared error (MSE).

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

where (m) is the number of data points, (y) is the observed value, and ($\hat{y}_i$) is the predicted value.

Train Model:

The model is trained on a dataset, where the algorithm learns the values of the coefficients that best fit the data. This involves feeding the algorithm input-output pairs and adjusting the coefficients until the model produces predictions close to the actual outcomes.

Linear regression relies on the assumption of a linear relationship between independent and dependent variables, normally distributed errors, constant error variance (homoscedasticity), and the absence of perfect multicollinearity, ensuring that there is no perfect linear relationship among the predictors.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)

Total Marks: 3 marks (Do not edit)

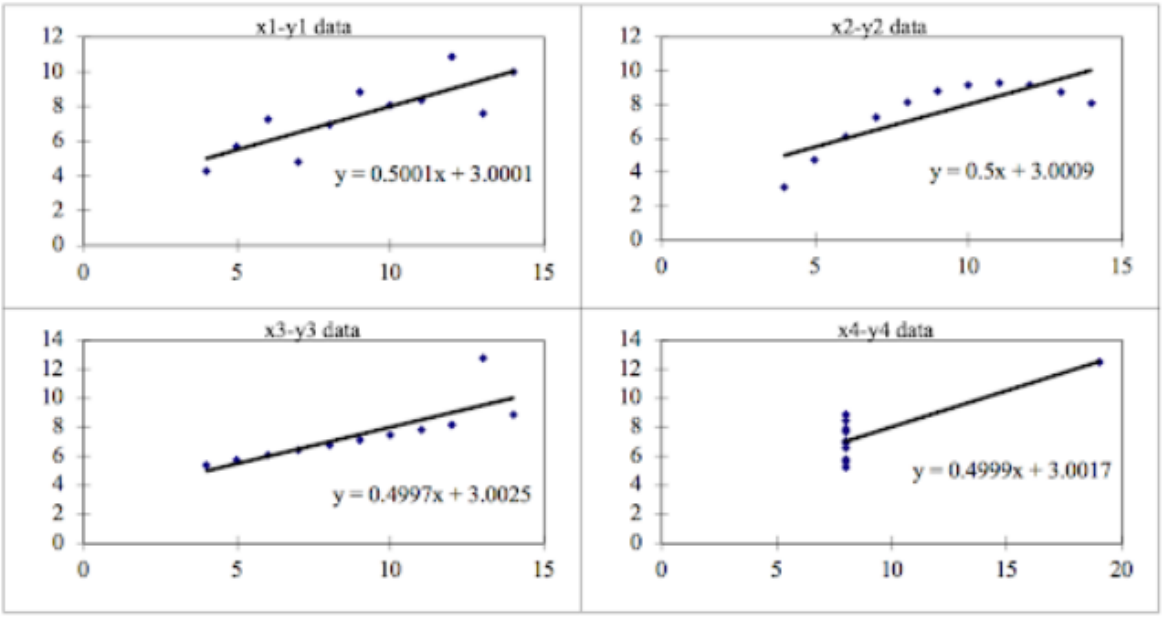
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet:

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

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

Based on the data above we can plot four different scatter plot



Anscombe's Quartet Four Datasets :

Data Set 1: fits the linear regression model pretty 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.

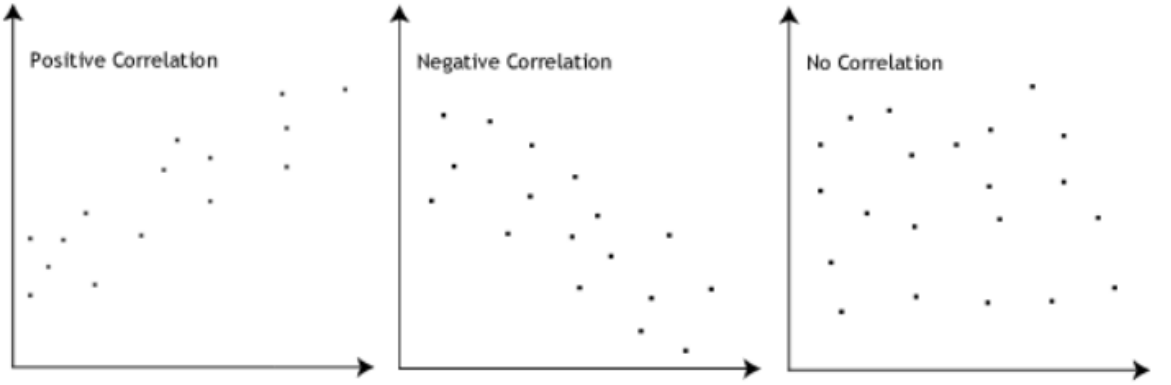
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

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)

- The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a 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).
- The Pearson correlation coefficient, *r*, 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.



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)

- In statistics, scaling is the process of transforming data to fit a specific scale, such as 0–1 or 0–100. It's used to make data more manageable and comparable, especially when working with variables that have different units or ranges.
- scale refers to the range or spread of the values in a dataset. We use scaling to make data more manageable and comparable, especially when dealing with variables that have different units or vastly different ranges

- Scaling ensures that all the variable in analysis are used properly.

Normalized scaling

- Also called min-max scaling, this technique scales feature values to a range between 0 and 1. It's best for data that doesn't follow a normal distribution, and is useful when the scale of features varies greatly. Normalization makes no assumptions about the underlying data distribution

Standardized Scaling

- Also called Z-score scaling or zero-mean scaling, this technique scales data values so that they have a mean of 0 and a standard deviation of 1. It's best for data that follows a normal distribution, and is often used when algorithms assume normality. Standardization doesn't always have a bounding range, so outliers in the data won't be impacted

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)

<Your answer for Question 10 goes here>

- When the value of VIF is infinite, it usually indicates perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in a regression model are perfectly correlated (linearly dependent) with other variables.
- In such cases:
    - There is redundant information - One variable can be expressed as a perfect linear combination of others.
    - Matrix Inversion Issues - In the computation of the VIF, there's an attempt to invert a matrix, and perfect multicollinearity leads to the matrix being singular (noninvertible)

- When the matrix is singular, it means that one or more variables can be predicted exactly from the others, and as a result, the computation of the VIF becomes problematic, leading to an infinite VIF value.

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)

<Your answer for Question 11 goes here>

A Q-Q plot, or quantile-quantile plot, is a graphical tool that compares two probability distributions by plotting their quantiles against each other. It's a powerful method for assessing if data is normally distributed, and is often used in linear regression to check the normality of residuals

Q-Q plot tells us about:

1.	**Distribution similarity**
	if the points on the plot fall approximately along the 45-degree reference line, the two distributions are similar
2.	**Distribution differences**
	If the points on the plot deviate from the reference line, the two distributions are different.
3.	**Distribution shape**
	The shape of the points on the plot indicates how the distributions compare in terms of location, scale, and skewness. For example, a J-shape indicates positive skewness.
4.	**Outliers**
	Outliers are indicated by points that are far from the general pattern of data points.

In linear regression, a Q-Q plot is used to assess the normality of residuals. The theoretical distribution is plotted on the x-axis, and the model residuals are plotted on the y-axis. If the residuals are normally distributed, the points on the plot will closely follow a straight line at a 45% angle upwards.