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)

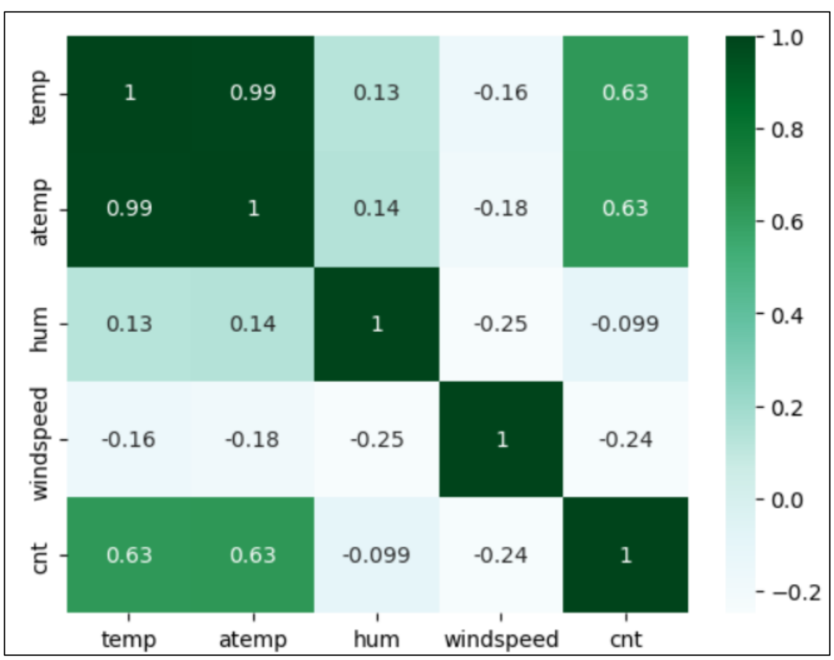
- Bike demand in the September is the highest.
- Bike demand takes a dip in spring.
- Bike demand in year 2019 is higher as compared to 2018.
- Bike demand is low in winters
- The demand of bike is high if temperature is more.
- Bike demand doesn’t change whether day is working day or not.

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 helps to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.
It also reduces the collinearity between dummy variables

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)

atemp and temp both have the highest correlation of 0.63 each, with 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)
To validate assumptions of the model, we check following things:

1. Linearity of relationship between target and predictor variables.
2. Residual Analysis: Check if error terms are normally distributed
3. Error terms are independent of each other

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)

Top 3 features significantly contributing towards demand of shared bikes are:
1) moth of September (with coefficient: 642.5944)
2) yr (with coefficient: 2049.0761)
3) spring season (negatively impacts the demand with coefficient: - 2686.6777)

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)

<Your answer for Question 6 goes here>
Linear regression algorithm is a form of regression, where the target variable is continuous. It estimates the relationship between a target variable and one or more predictor variables.

The Equation of linear Regression is y= m1x1 + m2x2 + m3x3 + ……….. + m(n) x(n) + c .
Where y is target variable and x1, x2 , x3 …… xn are predictor variables .
And we have two unknowns m, and c. We need to choose those values of m and c, which provide us with the minimum error.

We need to get the best fit line which is the line that has the minimum error.

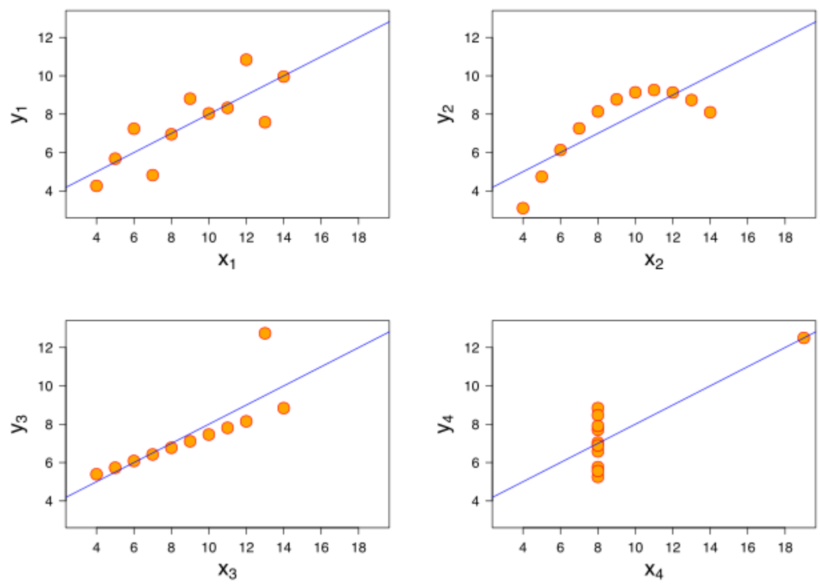
In linear regression, when the error is calculated using the sum of squared error, this type of regression is known as OLS, i.e., Ordinary Least Squared Error Regression.
Error function is explained by 'e = - y', and error depends on the values of 'm' and 'c'. Our aim is to build an algorithm which can minimize the error.
In order to do so we use cost function of Linear Regression, Which is: J(mi, c)= (1/2n)Σ(yi – yp)^2
Where yi and yp are expected values and predicted values.
Our main aim is to minimize J by changing m and c and it can be done using Gradient Descent Algorithm.

Cost function measures the performance of a Machine Learning model for given data.

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)

<Your answer for Question 7 goes here>

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed.



All four sets are identical when examined using simple summary statistics but vary considerably when graphed.

1) The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

2) The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

3) In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

4) the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

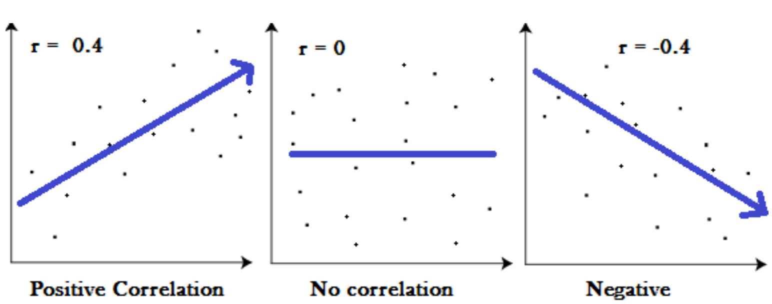
Anscombe's quartet

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

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)

<Your answer for Question 8 goes here>

Pearson's R or correlation coefficient is a measure of 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 relationship or correlation.



1) A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

2) A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.

3) Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.95| = .95, which has a stronger relationship than .55.

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)

<Your answer for Question 9 goes here>
When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. 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:

A. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

B. 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.

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>

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R^2 =1, which lead to $1/(1-R^2)$ infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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)

Q-Q (quantile-quantile) plots play a vital role in graphically analyzing and comparing two probability distributions by plotting their quantiles against each other. If the two distributions that we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight line y = x. A Q-Q plot tells us whether a data set is normally distributed.

We plot the theoretical quantiles, basically known as the standard normal variate (a normal distribution with mean of zero and a standard deviation of one) on the x-axis and the ordered values for the random variable, which we want to determine whether or not is a Gaussian distribution, on the y-axis. This gives a beautiful and smooth straight-line-like structure from each point plotted on the graph.

We have to focus on the ends of the straight line. If the points at the ends of the curve formed from the points are not falling on a straight line but are scattered significantly from these positions, then we cannot conclude a relationship between the x- and y-axes. This result clearly signifies that the ordered values that we wanted to calculate are not normally distributed.

If all the points plotted on the graph perfectly lie on a straight line, then we can clearly say that this distribution is normal because it is evenly aligned with the standard normal variate, which is the simple concept of Q-Q plot.

Q-Q plots are also used to find the skewness (a measure of asymmetry) of a distribution. When we plot theoretical quantiles on the x-axis and the sample quantiles whose distribution we want to know on the y-axis, then we see a very peculiar shape of a normally distributed Q-Q plot for skewness. If the bottom end of the Q-Q plot deviates from the straight line but the upper end does not, then we can clearly say that the distribution has a longer tail to its left. Put another way, it is left-skewed, also called negatively skewed. When we see the upper end of the Q-Q plot deviate from a straight line while the lower follows one, then the curve has a longer tail to its right and it is right-skewed, also called positively skewed.
