Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Inferences made from the categorical variables on their effect on our dependant variable 'cnt' are as below:

1. Rental bikes are in most demand during spring season followed by summer
2. Demand of rental bikes has increased significantly in 2019. So, clearly gaining popularity over the time.
3. There's an increasing demand for rental bikes till June. September has the highest demand. It then starts decreasing.
4. Renting of bikes doesn't seem to have much effect over the weekdays.
5. Holidays have decreased the demand a bit.
6. Rental bikes are in demand depending upon the weather. It has the most demand when the weather is clear or has very few clouds.

Q2. Why is it important to use drop_first = True during dummy variable creation?

Answer: This is because we only need n-1 dummy variables, where n = levels of categories in a variable. Even if drop one level, we are still able to explain all the n levels of a category. Drop_first = True helps us do that.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: temp and atemp variables have the highest correlation with the target variable cnt. We can see that on the pairplots for numerical variables.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: We concluded that our model is a best fit model as we have satisfied all 5 assumptions on the Linear Regression. Also calculated few statistical terms like R-squared values, F-statistic, p-values, RMSE, MAE and all these are within the acceptable range.

Assumptions on linear regression in detail.

1. There is linear relationship between our Independent and Dependant variables. we concluded that from the pairplots.
2. Error terms are normally distributed. we concluded that by plotting histogram of error terms

3. Error Terms have Homoscedasticity. we concluded that from the scatter plot.
4. There is no Auto correlation between variables. we concluded that from the Durbin Watson Statistic on our final model. If the value is 2 means no autocorrelation. Our final model 3 has a Durbin Watson score of 2.03 ~ 2
5. There are no Multicollinearity between our Independent variables. we concluded this from the VIF values from our final model. Our final model has VIF < 5

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: Top 3 Features contributing significant demands:

1. Yr (Positive correlation)
2. Temp (positive correlation)
3. Weathersit_3 (negative correlation)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer: Linear Regression is a statistical model that analyses dependant and independent variables which are continuous in nature. Mathematically, this can be represented by a line equation

$$y = mX + C$$

where y = Dependent variable, X = Independent, m = slope of the regression line, C = intercept / constant

Linear Regression models are divided into two types:

1. Simple Linear Regression – here we have only one independent variable
2. Multiple Linear Regression – here we have more than one independent variables

The strength of the linear regression model is explained by R^2 value. R^2 always takes values between 0 and 1. The higher the R^2 the better the model fits our data.

F-statistic is used to assess whether the overall model fit is significant or not. The higher the F-statistic value, the more significant a model turns out to be.

We determine our model is a best fit model based on R^2 score, F-statistic, p-values, and 5 important assumptions of linear regression are satisfied which are –

1. Linear Relationship between Independent and dependant variables
2. Error terms are normally distributed with mean 0.
3. Error terms have Homoscedasticity
4. No Autocorrelation between variables.
5. No Multicollinearity between independent variables

Q2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet shows us the importance of graphing the data and effect of outliers on the statistical properties by giving an example where 4 datasets have a similar simple descriptive statistic, yet they appear entirely different when graphed.

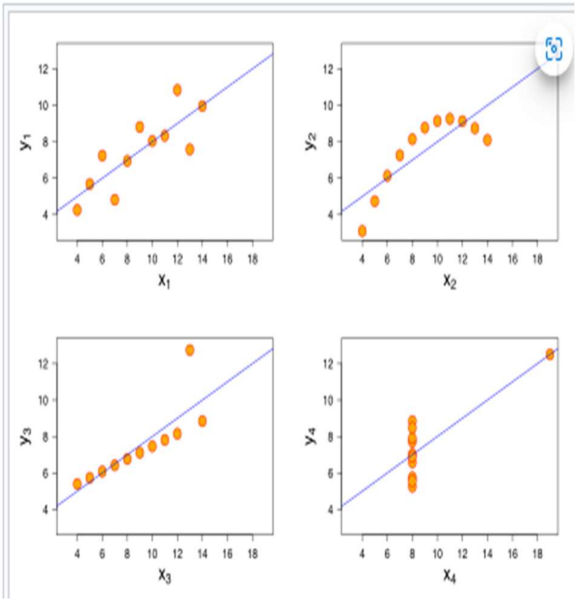
Anscombe's quartet comprises of four datasets which is shown as below. Each dataset consists of 11 (x,y) points. Now all these 4 datasets have almost similar descriptive statistics, also shown in fig, 1 below.

Now, when these datasets are graphed using scatter plot, see fig. 3 below. We can see how each dataset have a different tell.

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	$y = 3.00 + 0.500x$	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



- First scatter plot shows a simple linear relationship between x and y
- 2nd (top right) scatter plot shows clearly shows a non linear relationship between x and y
- 3rd (bottom left) plot shows a perfect linear relationship between all datapoints except one, which seems to be an outlier
- 4th one shows when one leverage point is enough to produce a high correlation coefficient even though other data points do not show linear relatioship

Q3. What is Pearson's R?

Answer: Pearson's R also known as Pearson correlation coefficient is a statistical test which measure the strength between different variables and their relationships.

Pearson's R returns a value between -1 and 1, where

-1 indicates a strong negative correlation between the variables.

0 indicates there's no relationship between the variables.

1 indicates a strong positive correlation between the variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

What? - Scaling is a step of data pre-processing which is applied to independent variables to handle highly varying magnitudes or values or units and normalize them within a particular range.

Why? - The idea of scaling is to bring highly varying magnitudes to within a particular range so that a machine learning algorithm understands it and helps speeding up the calculations in an algorithm. If scaling is not done then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. This could make our model unfit for interpretations.

Difference between normalized scaling and standardized scaling:

Normalization rescales our data into a range between 0 and 1 whereas Standardisation rescales our data into a standard normal distribution with mean 0 and standard deviation 1

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: If the value of VIF is infinite, it means there's a perfect correlation between the variables. A large VIF means the high linear correlation between independent variables. This happens due the presence of multicollinearity between independent variables. The higher the VIF, higher the multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: Q-Q plot is a short for Quantile – Quantile plot. This type of plot is used to determine whether a dataset follows a normal distribution or not. In linear regression this can be used to see if the error terms are normally distributed which happens to be one of the crucial assumptions of linear regression.