Assignment-based Subjective Answers

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

Sol: Here are some of the inferences I made from my analysis of categorical variables from the dataset on the dependent variable (Count)

1. Fall has the highest median, which is expected as weather conditions are most optimal to ride bike followed by summer.
2. Median bike rents are increasing year on as year 2019 has higher median then 2018, it might be due the fact that bike rentals are getting popular and people are becoming more aware about environment.
3. Overall spread in the month plot is reflection of season plot as fall months have higher median.
4. People rent more on non holidays compared to holidays, so reason might be they prefer to spend time with family and use personal vehicle instead of bike rentals.
5. Overall median across all days is same but spread for Saturday and Wednesday is bigger may be evident that those who have plans for Saturday might not rent bikes as it a non-working day.
6. Working and non-working days have almost the same median although spread is bigger for non-working days as people might have plans and do not want to rent bikes because of that.
7. Clear weather is most optimal for bike renting, as temperature is optimal, humidity is less, and temperature is less.

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

Sol: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Sol: By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

Sol: 1. Linear relationship between target and feature variables

2. Autocorelation in residuals

3. No Heteroskedasticity.
4. No Multicollinearity
5. Residuals must be normally distributed

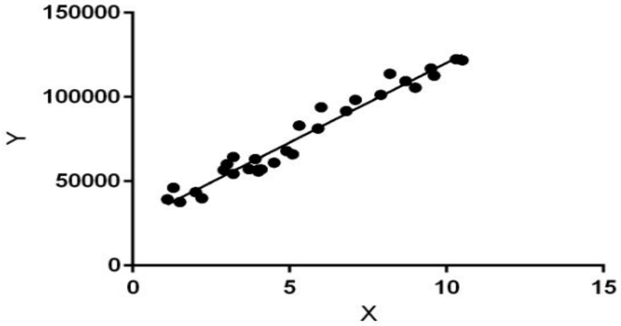
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Sol : The Top 3 features contributing significantly towards the demands of share bikes are:
weathersit_Light_Snow(negative correlation).
yr_2019(Positive correlation).
temp(Positive correlation).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Sol: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



2. ***Explain the Anscombe's quartet in detail.***

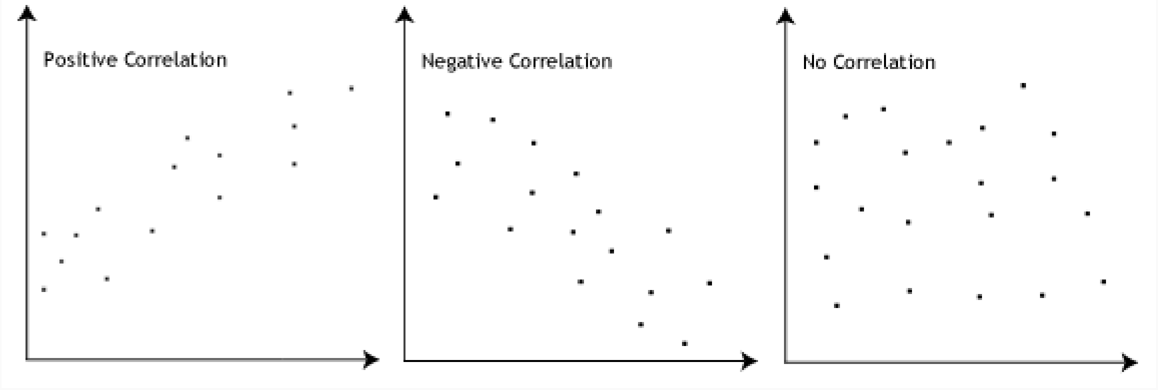
Sol: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. ***What is Pearson's R?***

Sol: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable
-

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

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	**Normalisation**	**Standardisation**
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-	It translates the data to the mean vector of

S.NO.	Normalisation	Standardisation
	dimensional data into an n-dimensional unit hypercube.	original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

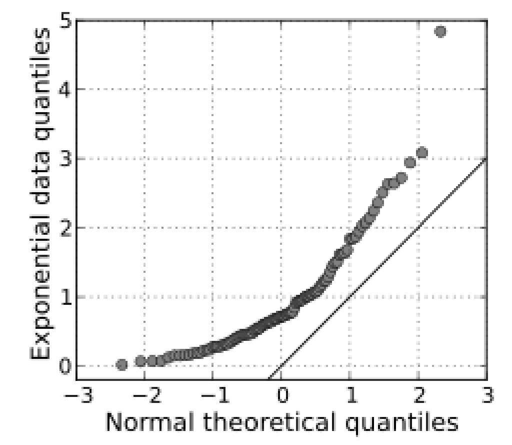
Sol: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multi co linearity.

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).

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

Sol: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.