Assignment-based Subjective Questions:

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

   Answer:

   #1. Season: 3:fall has highest demand for rental bikes
   #2. I see that demand for next year has grown
   #3. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing
   #4. When there is a holiday, demand has decreased.
   #5. Weekday is not giving clear picture about demand.
   #6. The clear weathershit has highest demand
   #7. During September, bike sharing is more. During the year end and beginning, it is less.

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

Answer:

drop_first=True was 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.

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

Answer:

Both temp and atemp 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?

Answer:

1. Comparisons of models theoretical calculations and results.
2. Comparisons of models coefficients and predictions with theory like p-value and VIF
3. Gathering and incorporating new data to check model prediction like various graph and plots.
4. Cross-validation/Data splitting.

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

Answer:
Temperature (temp) - A coefficient value of '0.2582' indicated that a unit increase in temp variable increases the bike hire numbers by 0.2582 units.

Month of May (mnth_may) - A coefficient value of '-0.0285' indicated that, w.r.t mnth_jan, a unit increase in mnth_may variable decreases the bike hire numbers by 0.0285 units.

Year (yr) - A coefficient value of '0.4958 ' indicated that a unit increase in yr variable increases the bike hire numbers by 0.4958 units.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

   Answer:

We train a model to predict the behaviour of your data based on some variables in linear regression. Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The two variables (independent and dependent) should be linearly correlated.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \implies y = a_0+a_1x$$

y= Dependent Variable.

x= Independent Variable.

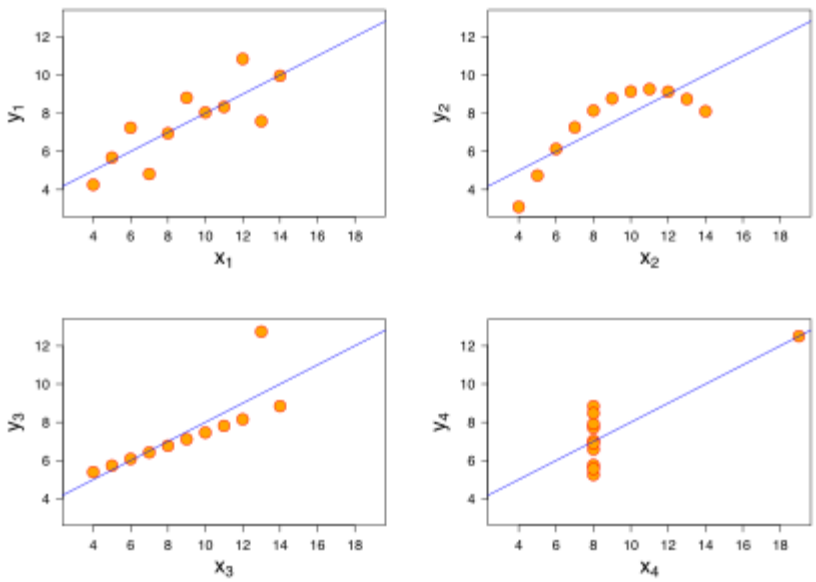
a0= intercept of the line.

a1 = Linear regression coefficient.

Example: Let's say we want to estimate the salary of an employee based on year of experience. You have the recent company data, which indicates that the relationship between experience and salary. Here year of experience is an independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.

2. Explain the Anscombe's quartet in detail.

Answer :
Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different in graph. Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.



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

For all four datasets:

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 :	0.67	to 2 decimal places

3. What is Pearson's R?

Answer:
Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient.

However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviation.

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

Answer:
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
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-squared, etc.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer:
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 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).

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

Answer:
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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 same distributions.