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: The categorical variables: ('mnth', 'weekday', 'season' & 'weathersit'), all of them affect the dependent variable ('cnt').

'season_2', 'season_4', 'mnth_9', 'weekday_1', 'weathersit_3' are the explicit categories that have major effect on the dependent variable, as they lower permissible value of VIF

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

Answer: The drop_first parameter specifies whether or not you want to drop the first category of the categorical variable you're encoding. By default, this is set to drop_first = False . This will cause get_dummies to create one dummy variable for every level of the input categorical variable.

drop_first reduces the number of columns needed for the model, it reduces the complexity and increases the computational speed of the model. Currently we deal with very less number of columns, but industry models may have over 1000 features. So in that case drop_first would significantly increase the speed of the model.

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

Answer: atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

Answer:

- Checking for the error terms normality using statistical tests
- From the correlation matrix, there is no multicollinearity.
- The error terms had a constant variance.

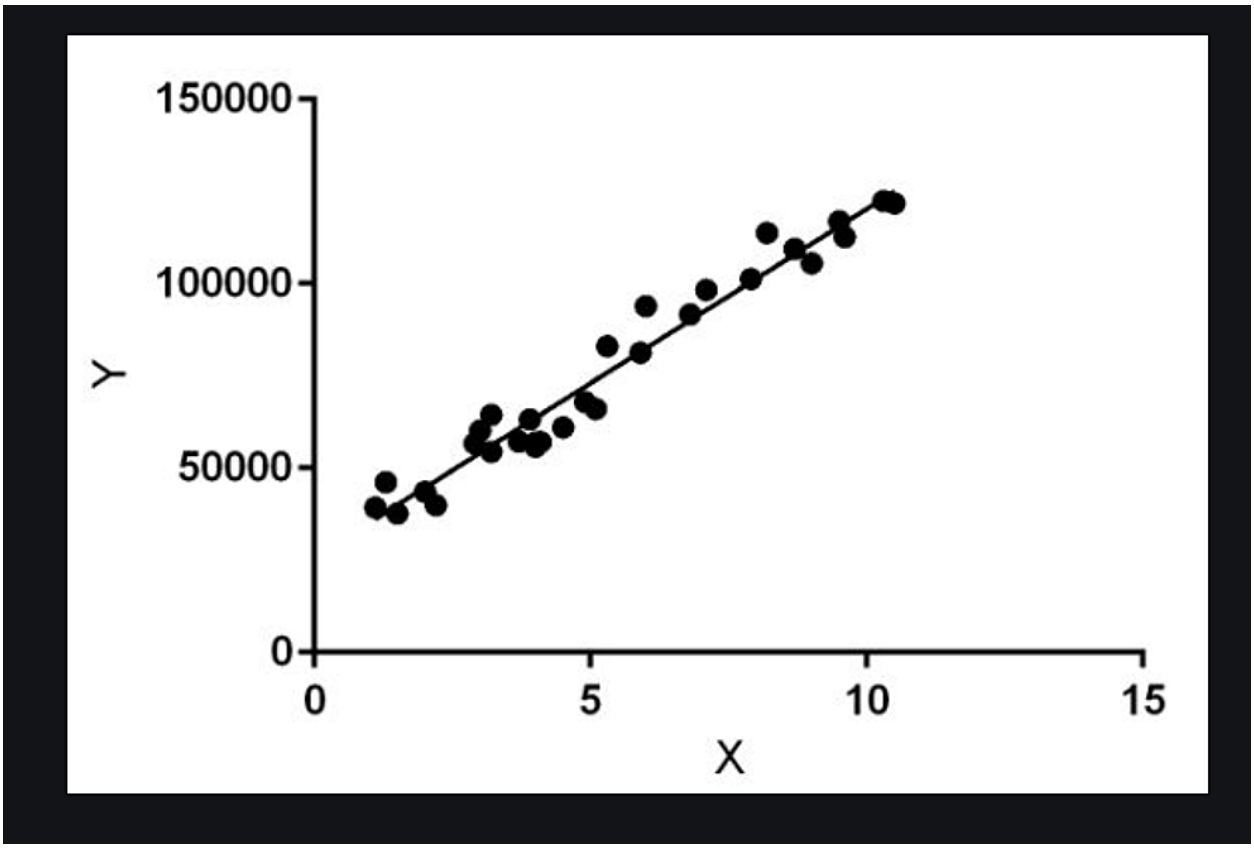
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: 'mnth', 'weekday', and 'weathersit'

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: 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.



Linear regression performs the task to predict a dependent variable value (y) based on a

given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$y = \theta 1 + \theta 2 \ x$

While training the model we are given :

x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x
Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have disparate distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

3. What is Pearson's R?
Answer: The Pearson correlation coefficient (r) is **the most common way of**

measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

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 time, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude into 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 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.

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).

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

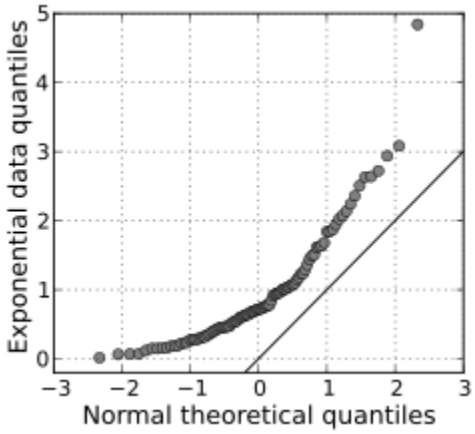
Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing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