Assignment-based Subjective Question

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

   I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualisation –

   - Fall season seems to have attracted more booking. And in each season the booking count has increased drastically from 2018 to 2019.
   - Most of the bookings has been done during the month of may, june, july, aug, sep and oct.Trend increased starting of the year tillmid of the year and then it started decreasing as we approached the end of year. Number of booking for each month seems to have increased from 2018 to 2019.
   - Clear weather attracted more booking which seems obvious. And in comparison to previous year, i.e 2018, booking increased for each weather situation in 2019.
   - Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
   - When its not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
   - Booking seemed to be almost equal either on working day or non-working day. But, the count increased from 2018 to 2019.
   - 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.
2. Why is it important to use drop_first=True during dummy variable creation?
   Answer:

   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.

   Syntax-

   drop_first: bool, defaut False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

   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 A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer:
   The numerical variable 'registered' has the highest correlation with the target variable 'cnt', if weconsider all the features.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Answer:
   I have validated the assumption of linear regression model based on below 5 assumption-

   - Normal of error terms
     - Error terms should be normally distributed
   - Multicollinearity check
     - There should be insignificant multicollinearity among variables.
   - Linear relationship validation
     - Linearity should be visible among variables.
   - Homoscedasticity
     - There should be no visible pattern in residual values.
   - Independence of residuals
     - No auto-correlation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Answer:
   As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (temp) - A coefficient value of '0.0936' indicated that a unit increase in temp variable increases the bike hire numbers by 0.0936 units.
- Year (yr) - A coefficient value of '0.2266' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2266 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2678' indicated that, a unit increase in weathersit_3 variable decreases the bike hire numbers by 0.2678 units.

So it's suggested to consider these variables utmost importance while planning, to achive maximum booking

The next best features that can also be considered are

- season_4: - A coefficient value of '0.1386' indicated a unit increase in season_4 variable increases the bike hire numbers by 0.1386 units.
- windspeed: - A coefficient value of '-0.1044' indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.1044 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   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.

$$Y = \theta_1 + \theta_2.x$$

Here, θ_1 is intercept

θ_2 is the coefficient of x

x: input training data

y: labels to data

Once we find the best θ_1 and θ_2 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.

We use cost function to update the value of θ_1 and θ_2 to get the best fit line.

Cost function(J) of linear regression is the root mean squared error(RMSE) between predicted y value (pred) and true y value(y).

2. Explain the Anscombe's quartet in detail. (3 marks)
   Answer:
   Anscombe's Quartet can be defined as a group of four data set 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.

Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

The four datasets can be described as:

1. Dataset 1**:** this fits the linear regression model pretty well.

2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R? (3 marks)
   Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Answer:
   Scaling 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 scalling to bring all the variables to the same level of magnitude.

It is important to note that scalling just affects the coefficients and none of the other parameters like t-statistics, F-ststistic, p-values, R-squared, etc.

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-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of 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.

S.NO.	Normalisation	Standardisation
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? (3 marks)
   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. (3 marks)
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.

The quantile-quantile(q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantile of the first data set against the quantiles of the second data set.

The slope tells us whether the steps in our data are too big or too small. For example, if we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes (a.w.a quantiles) compare between our data and the normal distribution.

A ateeply slopping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. One example cause of this would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).