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?

Created box plots to analyze categorical columns. By looking at the box plots we can infer the below points:

More bookings are done in the fall season followed by summer, winter, and spring

More bookings are done in the fall and summer months - May through sep (excluding July)

No drastically change whether its weekday or weekend

More bookings when weathersit is clear

No drastically change whether it working day or not

More booking in the year 2019 compared to 2018

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

If a given categorical feature has more than one level, we use dummy variable creation; so that we can transform the data into n-1 levels. Let's take an example of the season feature, as it has 4 levels (spring, summer, winter, and fall), we can transform and create 4 columns with values of 1 or 0. Out of the 4 columns that were created we can drop the first one and say when the other three columns are zero the first column (that is dropped) is true. drop_first=True is important, as it helps to reduce the extra column creation while doing dummy variable creation.

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

temp variable has a high correlation with y-axis (target variable).

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

Validate the assumptions of Linear Regression Model based on below assumptions:

Multicollinearity Check	- Multicollinearity between variables should be insignificant
Linear Relation validation	- Linearity should exist between the variables
Error Terms	- Normally distributed error terms
p-value	- Low p-value
VIF	- Low VIF

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

Please see below tops 3 features contributing significantly towards explaining the demand of the shared bikes

- temp
- year
- winter

General Subjective Questions:

----

1) Explain the linear regression algorithm in detail

Linear Regression is commonly used for predictive analysis and modeling. Linear Regression is a Machine learning algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable (target) based on given independent variables. So, this regression technique finds out a linear relationship between a dependent and the other given independent variables.

Use cases of Linear Regression - Prediction of trends and sales targets - To predict how industry is performing or how many sales targets industries may achieve in the future

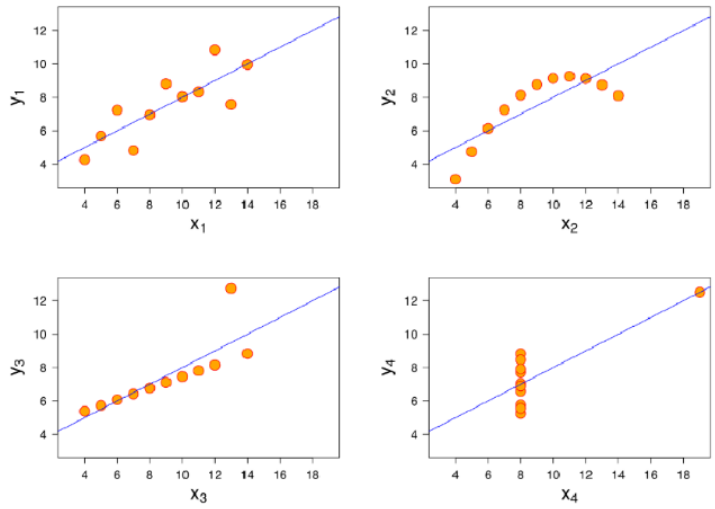
Mathematically the relationship can be represented with the help of the equation - Y = mX+c (Y is the dependent variable/we try to predict, X is the independent variable, m is slope and c is constant)

Linear Regression are of two types - Simple Linear Regression and Multiple Linear Regression

2) Explain the Anscombe's quartet in detail

Anscombe's Quartet(AQ) is nothing but, 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. AQ was developed by statistician Francis Anscombe. It comprises four datasets, each dataset consists of eleven (x,y) points. The essential thing to note about these datasets is that they share same descriptive statistics, but things can change completely when they are plotted.

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



By looking at the above plots we can say :

- Dataset 1 appears clean and well-fitting linear model
- Dataset 2 is not normally distributed
- Dataset 3 is linear, but there are outliers
- Dataset 4 shows that one outlier is enough to produce a high correlation coefficient

3) What is Pearson's R?

The Person correlation method is the most common method to use for numerical variables, it assigns a value between -1 and 1, where 0 is no correlation, 1 is a total positive correlation, and -1 is a total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Greater than 0.5 is Strong, between 0.3 and 0.5 is moderate, between 0 and 0.3 is weak, 0 is None, between 0 and -0.3 is weak, between -0.3 and -0.5 is moderate, and less than -0.5 is strong.

The formula used is –

Formula

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

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

Scaling is a technique to standardize the independent feature present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then machine learning tends to weigh greater values, and higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalized scaling – Min and Max values of features are used for scaling. It is used when features are of different scales. Scales values between (0,1) or (-1,1). It is really affected by outliers

Standardized scaling – Mean and standard deviation is used for scaling. It is used when we want to ensure zero mean and unit standard deviation. It is not bounded to a certain range. It is much less affected by outliers.

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

If there is a perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

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

In statistics, a Q–Q plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point on the plot corresponds to one of the quantiles of the second distribution plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity 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.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences.