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?

Ans.: -

There were 6 categorical variables in the dataset. I used Box plot (refer the fig above) to study their effect on the dependent variable ('cnt')

The inference that I could derive were:

1. **SEASON**: High demand bike booking were happening in fall with a median of over 5000 booking. This was followed by summer & winter with the total booking and very less demand in the spring season compared to others
   - Indicates:- season can be a good predictor for the dependent variable
2. **MNTH (MONTH)**: High demand bike booking were happening in the months june to sept with a median of over 4000 booking per month as compared to other months.
   - Indicates: mnth(month) has some trend for bookings and can be a good predictor for the dependent variable*
3. **WEATHERSIT** : High demand bike booking were happening during weathersit_clear_fewclouds with a median of close to 5000 booking. This was followed by weathersit_MIst_cloudy with the total booking.
   - Indicates:- weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
4. **HOLIDAY**: The bike booking were happening when it is not a holiday which means this data is clearly biased.
   - Indicates:- holiday CANNOT be a good predictor for the dependent variable.
5. **WEEKDAY**: The demand of the bike booking shows very close trend having their independent medians between 4000 to 5000.
   - This variable can have some or no influence towards the predictor. Let the model will decide if this needs to be added or not.
6. **WORKINGDAY**: There is no significant change in bike demand with workign day and non working day.The bike booking were happening in 'workingday' with a median of close to 5000 booking.
   - Indicates, workingday can be a good predictor for the dependent variable

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

Ans.: -

A **dummy** variable is a binary variable that indicates whether a separate categorical variable takes on a specific value. For a categorical variable that takes on more than one value, it is useful to create one **dummy** variable for each unique value that the categorical variable takes on.

The get_dummies() **function** is used to convert **categorical variable** into **dummy**/indicator **variables**. Data of which to get **dummy** indicators. String to append **DataFrame** column names. If appending prefix, separator/delimiter to use.

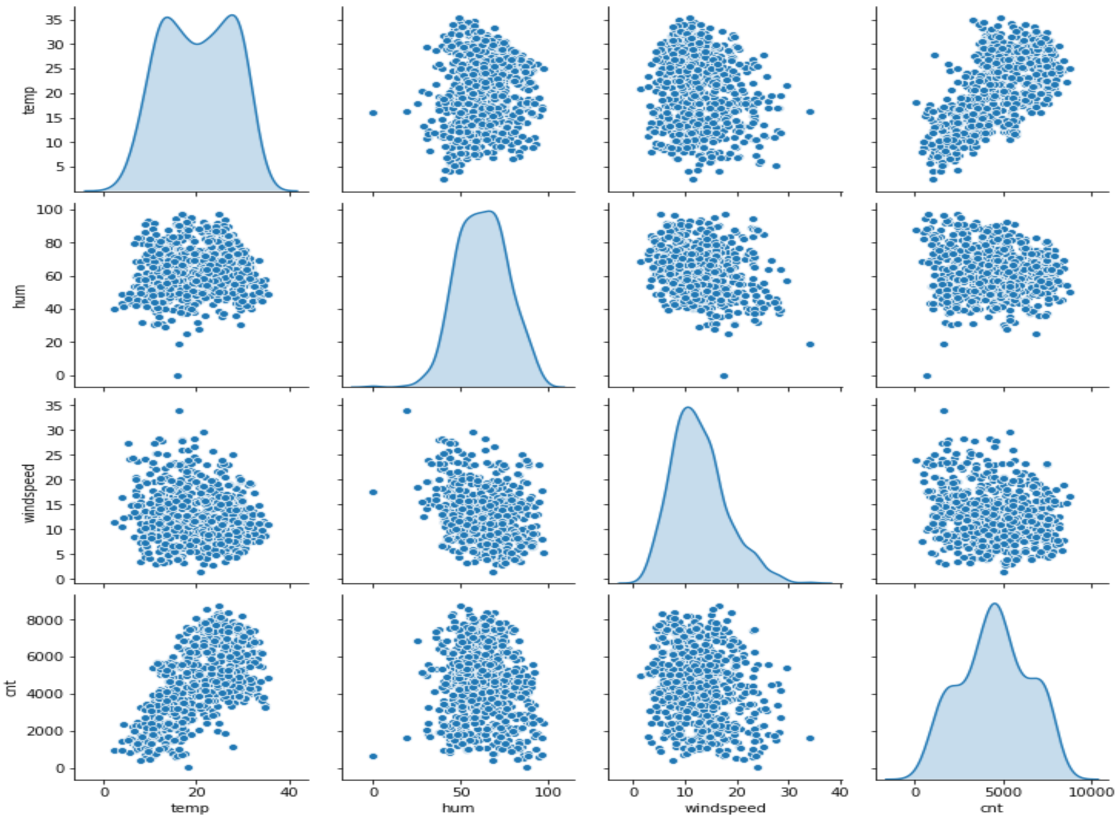
In the simplest case, we would **use** a 0,1 **dummy variable** where a person is given a value of 0 if they are in the control group or a 1 if they are in the treated group. **Dummy variables** are useful because they enable us to **use** a single regression equation to represent multiple groups. In our **DataFrame** we have "season, mnth, weekday and weathersit columns that columns we created a dummys for batter operation. We can perform groupby operation to get batter analysis visualization.

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

Ans.: -

Pair-Plot tells us that there is a LINEAR RELATION between **'temp'** and **'cnt'** variables has good relationship. We can say that **'temp'** and **'cnt'** variables we can consider while performing analysis.

This Scatter plot has **'temp', 'hum', 'windspeed'** and **'cnt'** variables and all are paired with each other. For reference I used an image which I taken from our **BookBike** data.



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

Ans.: -

Linear regression is probably the most important model in model building. Despite its apparent simplicity, it relies however on a few key assumptions (linearity, homoscedasticity, absence of multicollinearity, independence and normality of errors). Good knowledge of these is crucial to create and improve your model.

Because **we** are **fitting a linear model**, **we assume** that the relationship really is **linear**, and that the errors, or residuals, are simply random fluctuations around the true line. **We assume** that the variability in the response doesn't increase as the value of the predictor increases.

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

Ans.: -

In final model we are using these top 10 variables. I have taken top 15 variables for model building in which I have dropped 5 variables having high P-value and VIF.

As per our final Model, the top 3 features contributing significantly towards explaining the demand of the shared bike

Temperature (temp)
A coefficient value of '0.555' indicated that a unit increase in temp variable increases the bike hire numbers .
Light_Snow + Rain
A coefficient value of '-0.280' indicated that, w.r.t Weathersit_Mist_cloudy, a unit increase in Weathersit_LightSnow_LightRain variable decreases the bike hire numbers.
Year (yr)
A coefficient value of '0.241' indicated that a unit increase in yr variable increases the bike hire numbers

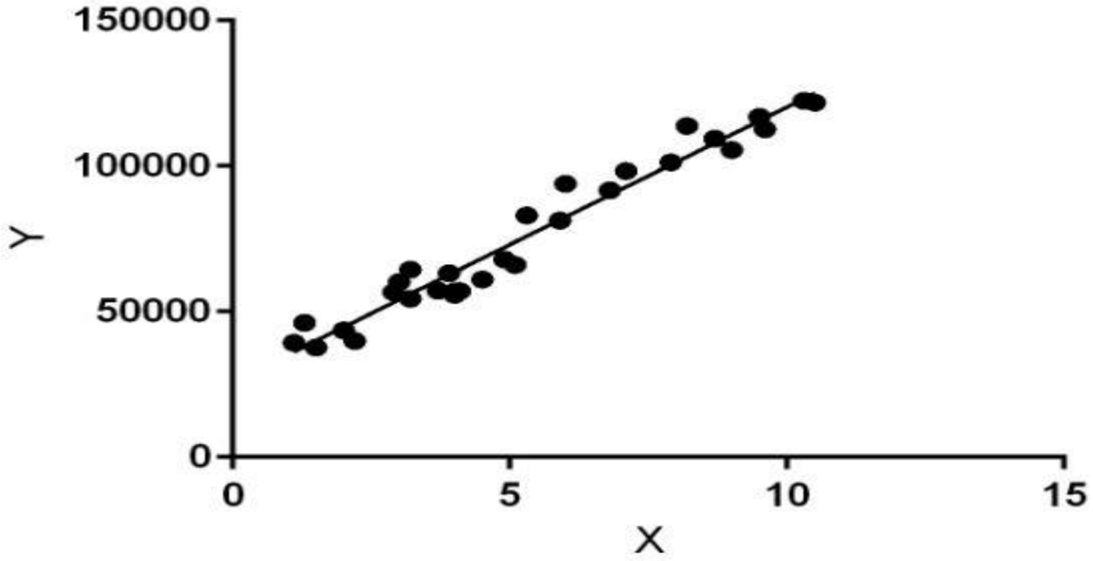
SO IT IS RECOMMENDED TO GIVE THESE VARIABLES UTMOST IMPORTANCE WHILE PLANNING, TO ACHIEVE MAXIMUM DEBOOKING

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans.: -

Linear Regression algorithm is a **machine learning algorithm** based on **supervised learning method**. 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 being used.



(Above image shows X as independent variable and Y is dependent variable)

Linear regression quantifies the relationship between one or more predictor variable(s) and one outcome variable. For **example**, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line.

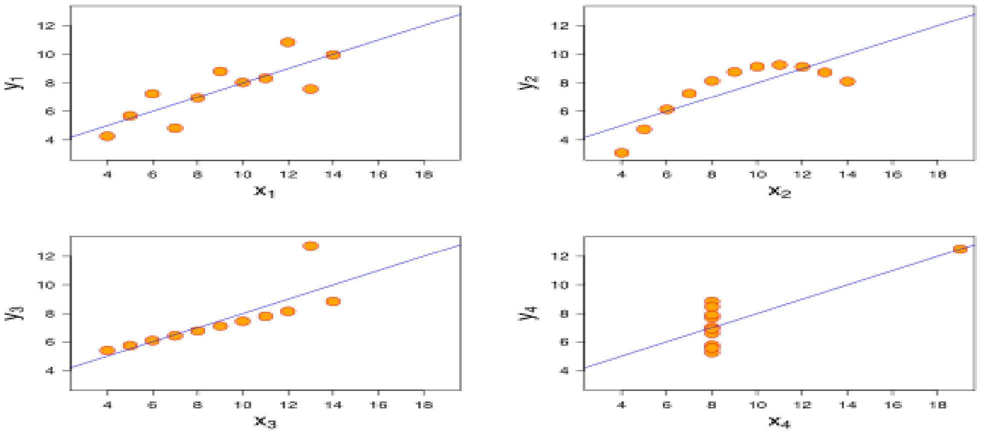
Linear regression is a way to model the relationship between two variables. ... The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

Linear regression attempts to model the relationship between two variables by fitting a **linear** equation (= a straight line) to the observed data. One variable is considered to be an explanatory variable (e.g. your income), and the other is considered to be a dependent variable (e.g. your expenses).

2. Explain the Anscombe's quartet in detail.

Ans.: -

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. 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 analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.



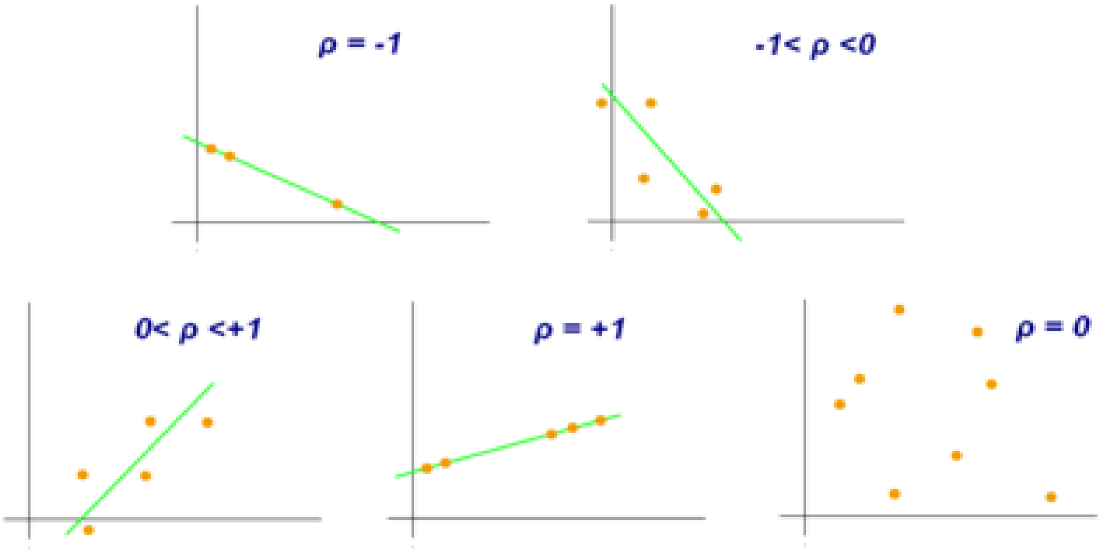
(All four sets are identical when examined using simple summary statistics, but vary considerably when graphed**)**

3. What is Pearson’s R?

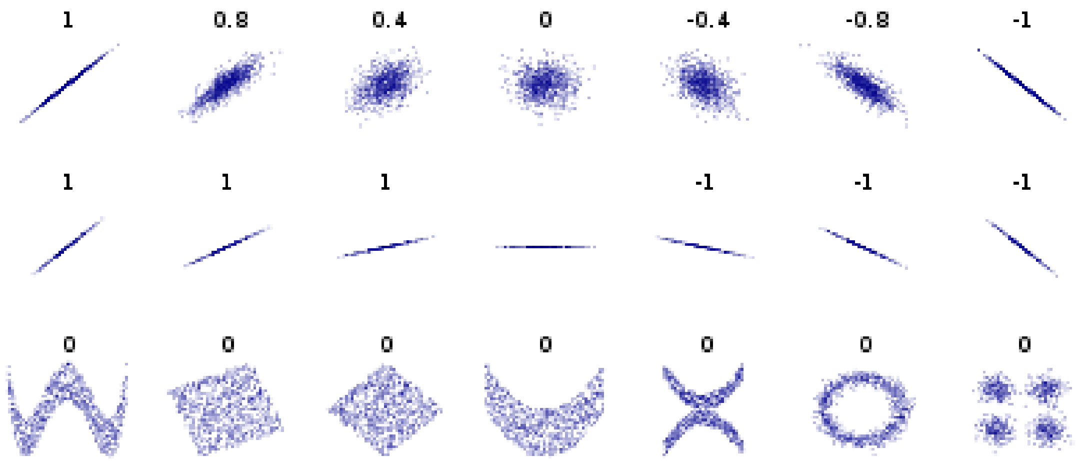
Ans.: -

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 statistic that measures linear correlation between two variables *X* and *Y*. It has a value between -1 and +1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.



(Examples of scatter diagrams with different values of correlation coefficient (ρ))



Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

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

Ans.: -

There are four levels of measurements: nominal, ordinal, interval, and ratio. The measurement scales, commonly used in marketing research, can be divided into two **types**; comparative and non-comparative scales. There is no unique way that you can use to select a particular **scaling** technique for your research study

Feature **scaling** is a method used to normalize the range of independent variables or features of **data**. In **data** processing, it is also known as **data** normalization and is generally performed during the **data** preprocessing step.

Feature Scaling is a technique to standardize the independent features 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 a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

The terms *normalization* and *standardization* are sometimes used interchangeably, but they usually refer to different things. *Normalization* usually means to scale a variable to have a values between 0 and 1, while *standardization* transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a **z-score**, and data points can be standardized with the following formula:

$$z_i = \frac{x_i - \bar{x}}{s}$$

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

Ans.: -

VIF stands for Variance Inflation Factor. During regression analysis, **VIF** assesses whether factors are correlated to each other (multicollinearity), which could affect p-values and the model isn't going to be as reliable.

There are some guidelines we can use to determine whether our VIFs are in an acceptable range. A rule of thumb commonly used in practice is if a **VIF** is > 10, you have high multicollinearity. In our case, with values around 1, we are in **good** shape, and can proceed with our regression.

If the **VIF** is equal to 1 there is no multicollinearity among factors, but **if** the **VIF** is greater than 1, the predictors may be moderately correlated. A **VIF** between 5 and 10 indicates **high** correlation that may be problematic.

The **Variance Inflation Factor** (**VIF**) is a measure of colinearity among predictor variables within a multiple regression. It is **calculated** by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone.

As the calculation of the **VIF** can be quite time consuming, you may choose to use only a random sample of 1000 pixels to calculate the **VIF**. 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).

As the calculation of the VIF can be quite time consuming, you may choose to use only a random sample of 1000 pixels to calculate the VIF. This increases the speed of calculation considerably, even though the accuracy of the VIF values is degraded. However this should be sufficient to get a rough overview.

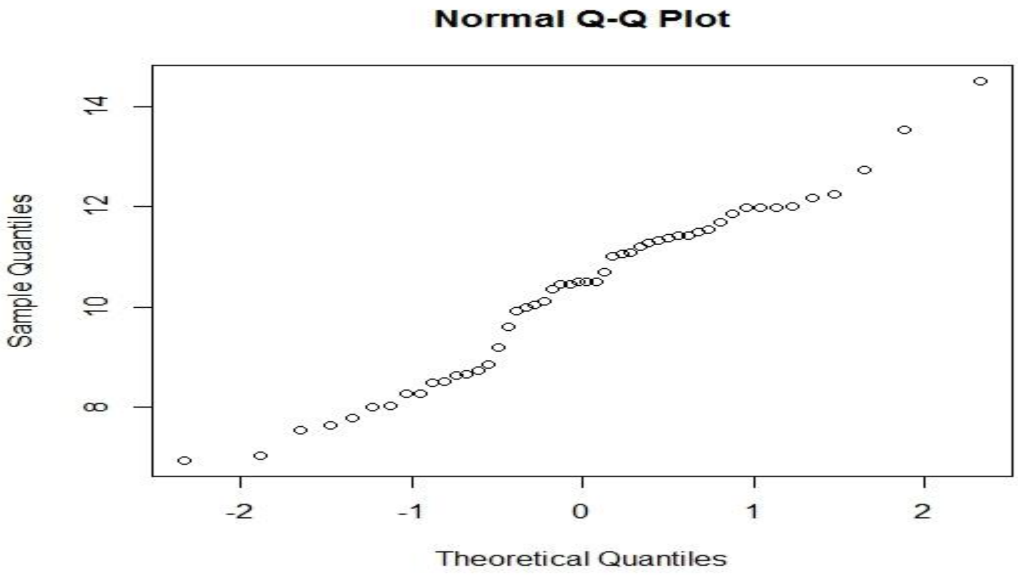
Try one of these to deal with VIF:

1. Remove highly correlated predictors from the model. If you have two or more factors with a high VIF, remove one from the model. ...
2. Use Partial Least Squares Regression (PLS) or Principal Components Analysis, regression methods that cut the number of predictors to a smaller set of uncorrelated components.

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

Ans.: -

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.



A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.