Q1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans)

Box plot and bar plot analysis has been done on categorical variable to understand the effect on dependent variable.

Below are the observations:

- season: Almost 32% of the bike booking were happening in Fall with a median of over 5000 booking (for the period of 2 years). This was followed by summer & winter with 27% & 25% of total booking. This shows that season can be a good predictor for the dependent variable.
- mnth: Almost 10% of the bike booking were happening in the months from may to sep with a median of over 4000 booking per month. This indicates, mnth has some trend with no of bookings and hence can be a good predictor. By the end of the year, the bookings decrease
- weathersit: Almost 67% of the bike booking were happening during 'Clear' weather with a median of close to 5000 booking (for the period of 2 years). This was followed by Misty weather with 30% of total booking. This indicates, weathersit has some trend with bookings and hence can be a good predictor
- holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. Therefore, holiday CANNOT be a good predictor for the dependent variable.
- weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. We will let the model decide if this needs to be added or not.
- workingday: Almost 69% of the bike booking were happening when it was a working day with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable
- There is drastic increase in the counts of booking from year 2018 to 2019. We can expect that there can be a progressive increase in booking counts year by year.

Q2. **Why is it important to use drop_first=True during dummy variable creation?**

Ans)

It removes the first column which is created for the first unique value of a column. Hence it reduces the correlations created among dummy variables.

Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k

categorical levels by removing the first level.

Let us say we have 3 types of values (say A, B and C) 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. Hence while creating dummy, we add drop_first = True so that we can eliminate the redundancy.

Q3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Ans)

'temp' variable has the highest correlation with the target variable.

Q4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans)

I have validated the assumption of Linear Regression Model based on below 5 assumptions

- Normality of error terms
  Error terms should be normally distributed
- Multicollinearity check
  There should be insignificant multicollinearity among variables. VIF should be less than 5
- Linear relationship validation
  Linearity should be visible among variables
- Homoscedasticity
  There should be no visible pattern in residual values.
- Independence of residuals
  No autocorrelation

Q5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Ans)

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- Year
- Temperature
- winter

Q1. **Explain the linear regression algorithm in detail.**

Ans)

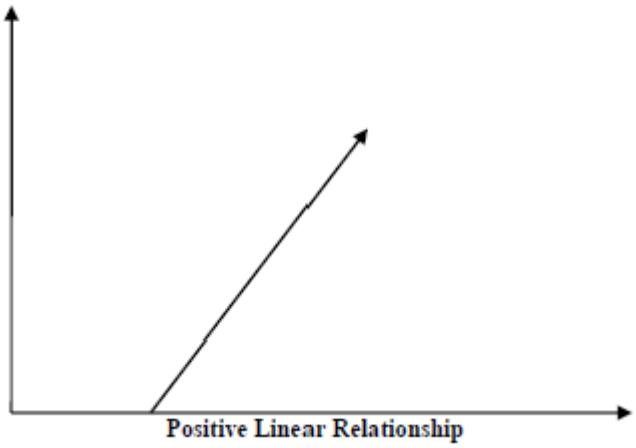
Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly
(increase or decrease).

Mathematically the relationship can be represented with the help of following equation −
Y = mX + c

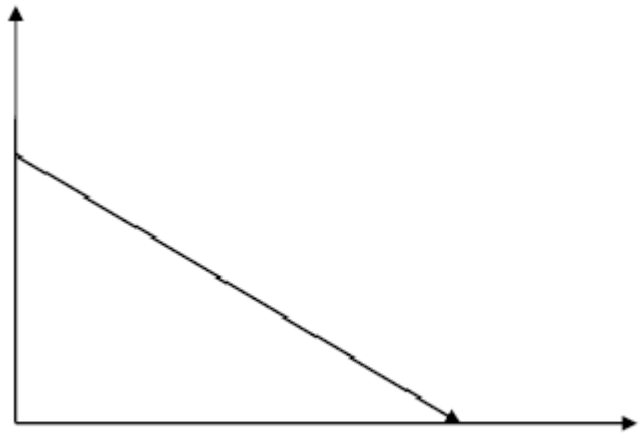
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship:
  A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph



- Negative Linear relationship:
  A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph



Negative Linear Relationship

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model −

- Multi-collinearity –
  Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

- Auto-correlation –
  Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

- Relationship between variables –
  Linear regression model assumes that the relationship between response and feature variables must be linear.

- Normality of error terms –
  Error terms should be normally distributed

- Homoscedasticity –
  There should be no visible pattern in residual values.

Q2. Explain the Anscombe's quartet in detail.

Ans)

The word quartet means a group or set of four. Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics. It appears very different when graphed.

There are eleven (x,y) points per data set.

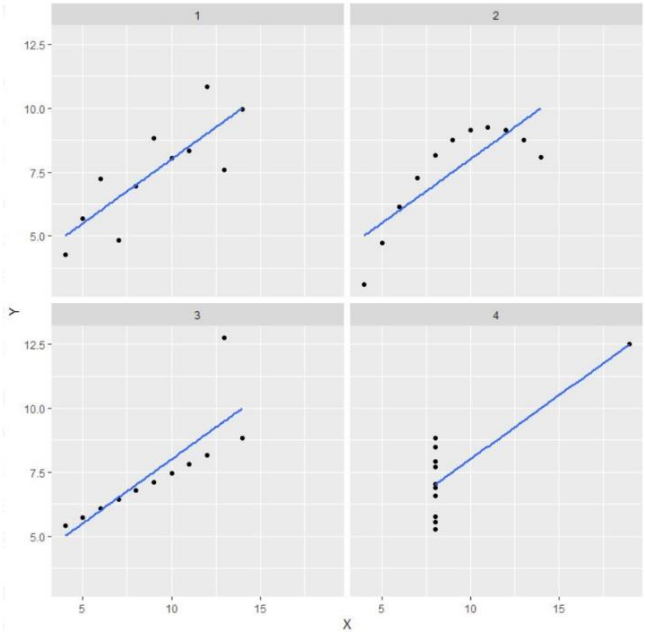
They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

I took these below screenshots from the internet only for explanatory purpose.

In this below data set, the mean, sd, correlation values of all the four data sets are almost similar.

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817



As mentioned above, it is very true that it appear very different when graphed.

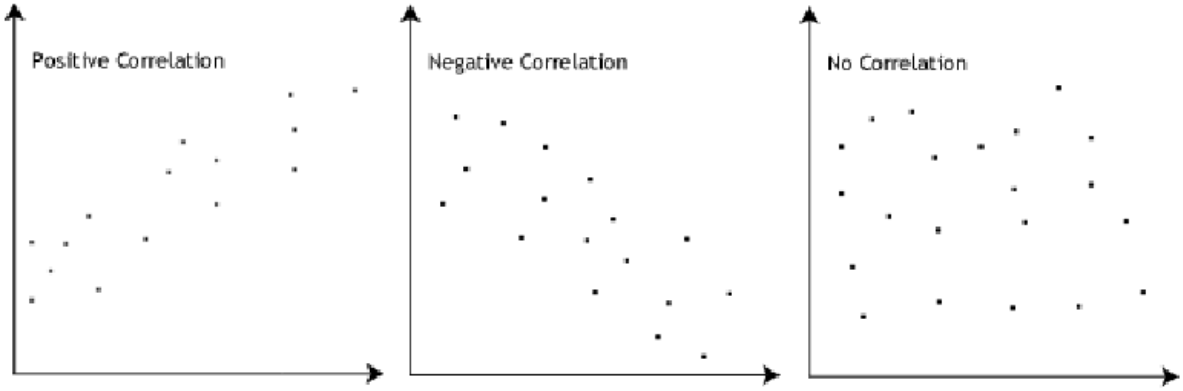
1) First graph tells us that the scatter plot seems to be have a linear relationship between x and y.
2) Second graph tells us that the relationship between x and y are non-linear.
3) Third graph (bottom left) tells us that there is a perfect linear relationship between all the datasets except one which seems like an outlier which indicates far from the line.
4) Fourth graphs explains that when one high-leverage point is enough to produce a high correlation coefficient.

Q3. **What is Pearson's R?**

Ans)

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.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.



Q4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Ans)

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.
Example: If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.NO.	Normalized scaling	Standardized scaling
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.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans)

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

If there is 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.

When the value of VIF is infinite, it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Uses:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance:

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. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests