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 1: Below is the inference from the analysis of the categorical variables and the effect on the dependent variable:

season: Almost 30% - 35% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, 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. I will let the model decide if this needs to be added or not.

workingday: Almost 69% of the bike booking were happening in 'workingday' 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

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

Ans 2: To analyze the categorical variable's effect on the dependent or target variable it needs to have only n-1 dummy variables, where n is the number of categories in that categorical variable. So drop_frist=True helps in reducing extra column creation during the dummy variable creation. It also reduces the correlation created among the dummy variables.

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

Ans 3: Looking at the pair-plot, temp and atemp variables have high correlation with the target variable. Between temp and atemp, it seems temp has little higher correlation with the target variable.

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

Ans 4: I validated below assumptions:

1. **Error terms are normally distributed with mean zero (not X, Y)**
   Validated it with the distribution plot.

2. **There is a linear relationship between X and Y**
   Validated this assumption with the pair plot.

3. **There is No Multicollinearity between the predictor variables**
   Validated the assumption with the VIF results of variables participating in the final model.

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

Ans 5: Temperature(temp), Weather Situation 3 (weathersit_3) and Year (yr) are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

Below is the reasoning:
Temperature (temp): A coefficient value of '0.5211' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5211 units.
Weather Situation 3 (weathersit_3): A coefficient value of '-0.2786' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by '0.2786' units.

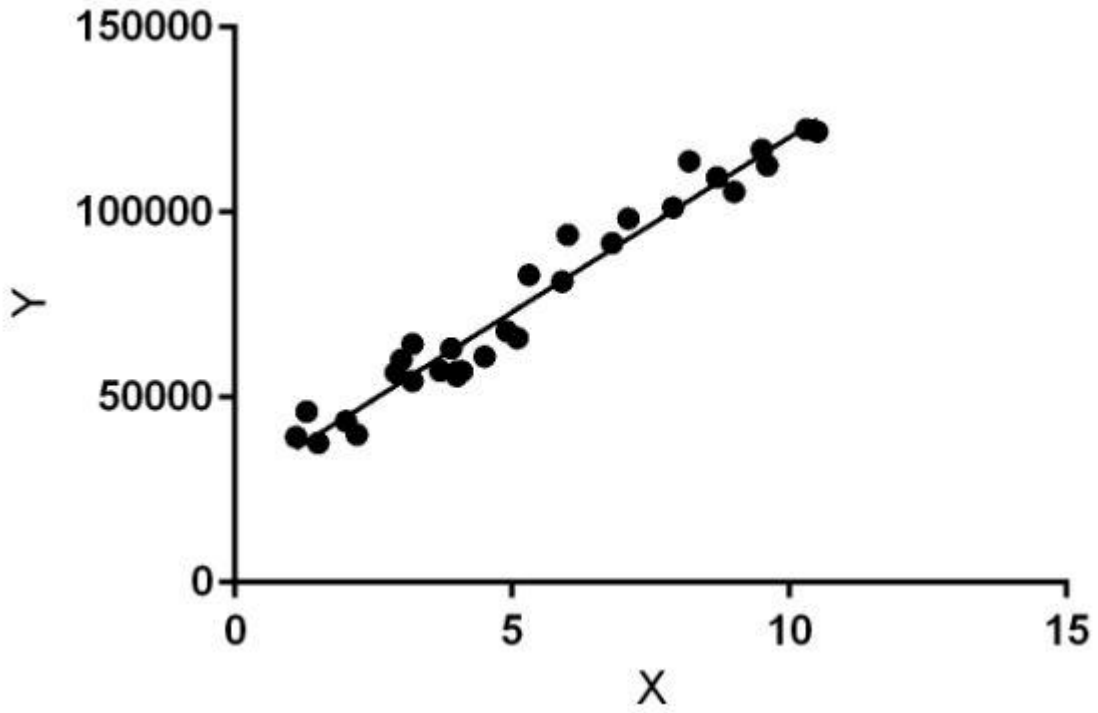
Year (yr): A coefficient value of '0.2328 indicated that a unit increase in yr variable, increases the bike hire numbers by 0.2328 units.

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail.

Ans 1: Linear Regression is a supervised machine learning method/algorithm that computes the linear relationship between the dependent variable and one or more dependent variables/features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and

independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).
Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings or to predict the future value of a currency based on its past performance.



One of the most important supervised learning tasks is regression. In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.
Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). 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.

Q 2. Explain the Anscombe's quartet in detail.
Ans 2:
Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were

created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

The four datasets of **Anscombe's quartet.**

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

Q 3: What is Pearson's R?

Ans 3: The Pearson correlation coefficient(r) also know as Pearson's 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.

Pearson's correlation is utilized when you have two quantitative variables and you wish to see if there is a linear relationship between those variables. Your research hypothesis would represent that by stating that one score affects the other in a certain way. The correlation is affected by the size and sign of the r.

You can interpret the value of 'r' as below:

- If r = -1, then there is a perfect negative linear relationship between x and y.
- If r = 1, then there is a perfect positive linear relationship between x and y.
- If r = 0, then there is no linear relationship between x and y.

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

Ans 4:

What

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.

Why

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 scaling to bring all the variables to the same level of magnitude.

It is important to note that 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.

  MinMax Scaling: x = x - min(x) / max(x) - min(x)

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).

Standardisation: x = x - mean(x) / sd(x)

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans 5: when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

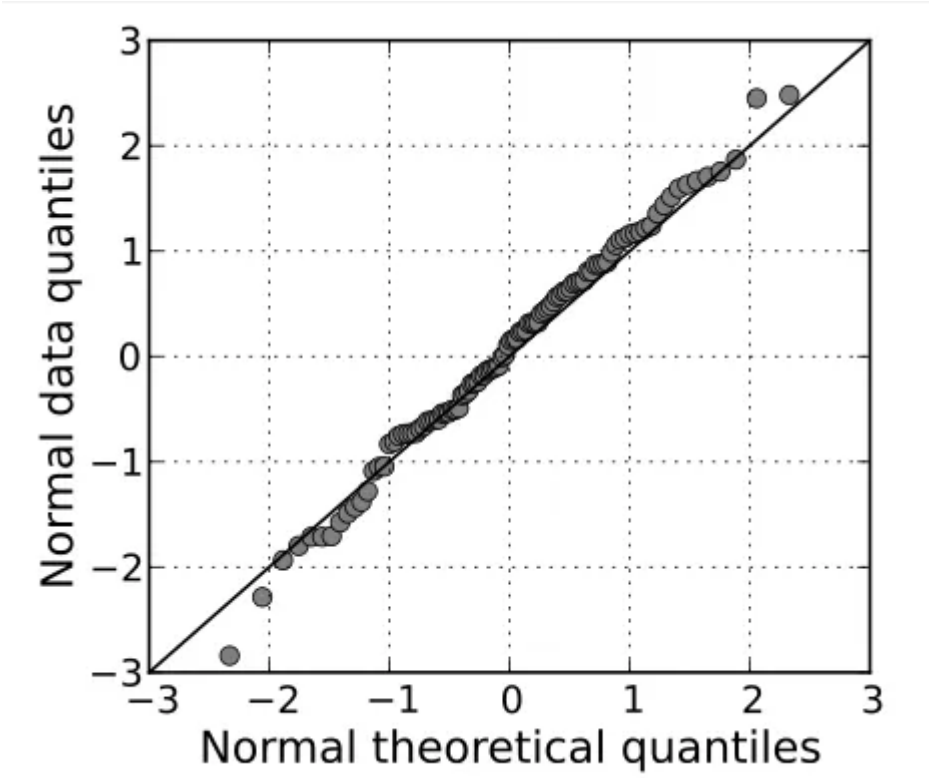
Q 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans 6: A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

In Statistics, Q-Q(quantile-quantile) plots play a very vital role in graphically analyzing and comparing two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

Being a data scientist or in general a statistician, it's very important for you to know whether the distribution is normal or not so as to apply various statistical measures on the data and interpret it in much more human-understandable visualization and there Q-Q plot comes into the picture. The most fundamental question answered by Q-Q plot is:

"Is the curve Normally Distributed?



Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. In general, we are talking about Normal distributions only because we have a very beautiful concept of 68–95–99.7 rule which perfectly fits into the normal distribution So we know how much of the data lies in the range of first standard deviation, second standard deviation and third standard deviation from the mean. So knowing if a distribution is Normal opens up new doors for us to experiment with the data easily. Secondly, Normal Distributions occur very frequently in most of the natural events which have a vast scope.

How does it work?

We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1)on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed…