Assignment-based Subjective Questions

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

After performing exploratory data analysis on the categorical variables, the following can be inferred:

i. Amongst the seasons, spring season has lesser bookings in the interquartile range, and the fall season has slightly higher median as compared to other seasons. It can be inferred that the season spring might have an effect on the number of bike rental bookings.
ii. From the month of January to September the median value of rental bookings increase, and from October to December they decrease.
iii. During the holidays, the median value of number of bike rental bookings reduces, with 75 percentile remaining similar and 25 percentile reducing. This means that during holidays, usually slightly less people rent bikes, on some days of holidays very few people booking the bikes.
iv. During snowy weather situations, the number of people bookings rental bikes is very few as compared to clear or misty weather.
v. The median value on all days of the week is almost the same, so this variable might not have an effect in the final model.
vi. On a working day, the median bike rental is similar to non-working day, but the 25th percentile is lower for weekends or holiday. It means that on weekends or holidays, there might be lower peaking in the number of rentals sometimes.

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

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. It gives k-1 dummies out of k categorical levels by removing the first level.

pandas.get_dummies(data, prefix=None, prefix_sep='_', dummy_na=False, columns=None, sparse=False, **drop_first=True**, dtype=None)

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

Looking at the pair plot, it seems that the variable ‘temp’ and ‘atemp’ have the highest correlation with the target variable.

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

1. By assuming a linear correlation between the parameters, the model thus obtained is able to predict the variation in target variable almost 80% . Thus linear regression is justified.
2. The error terms were normally distributed.
3. The error terms were independent of each other: there is no visible pattern in the residues.
4. Ensuring that the variance in error terms is constant i.e. homoscedacity. This was validated visually by plotting the predicted y value with given y value and comparing it with the regression line (ypred=ytest)

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

The top 3 features which affect the demand of shared bikes are temp, weathersit_snowy and yr i.e. temperature on the day, if the weather situation is snowy or not and if the year is 2018 or 2019.

General Subjective Questions

Explain the linear regression algorithm in detail

Linear regression is a type of supervised machine learning algorithm, which computes a linear relation between the target variable and the predictor variable(s).

$$y = b0 + b1 * x1 + b2 * x2 \ldots + bn * xn$$

Where Y is the target variable, x1,x2,..xn are predictor variables and b0,b1..bn are the constants

b0 is the intercept parameter of the line whereas b1,b2..bn are the slope parameters.

When only one predictor variable is present, it is called simple linear regression, else it is called multiple linear regression.

The linear regression algorithm is founded on reducing the ordinary least squared value (OLS algorithm)

$ei = yi - b0 - b1 * xi$ for simple linear regression, where ei is the error in the ith term

Residual sum of squares (RSS) is the sum of the squares of all the values of ei for a particular data set.

$$RSS = \sum_{1}^{N} ei^2$$

In OLS algorithm, RSS is the cost function which is to be minimized.

The following conditions should be met in any linear model:

i. Error terms should be normally distributed
ii. Error terms should be independent of each other
iii. Error terms should have a constant variance throughout the data set (homoscedacity)

Value of R^2 should be high.

$$R^2 = 1 - \frac{RSS}{TSS}$$

where ,

$$TSS = \sum_{1}^{n} (yi - ymean)^2$$

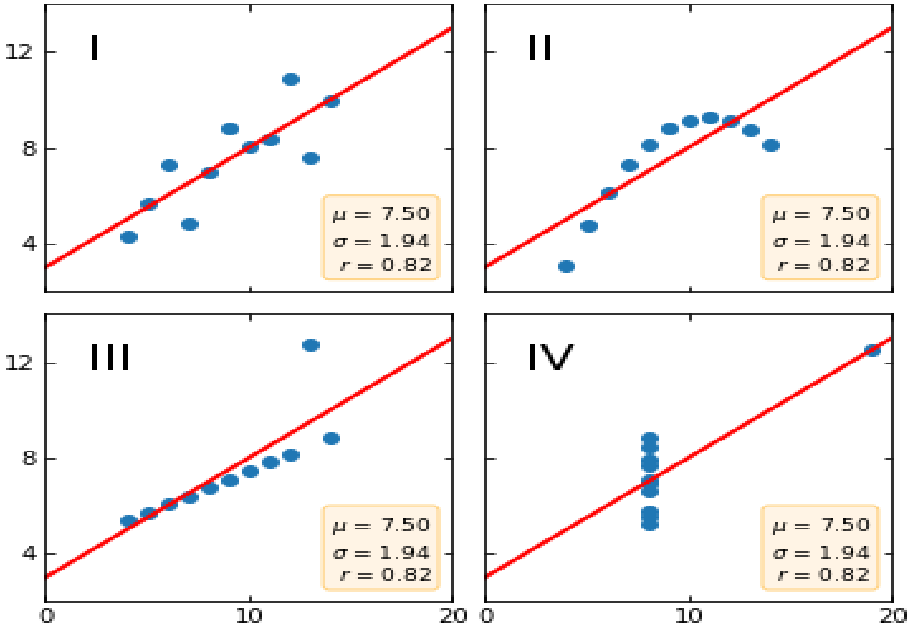
R^2 gives an estimation of the closeness of the model with the actual values, while also explaining the variance in the model.

In case of multiple linear regression, multicolinearity needs to be resolved as well. Multicolinearity means a model which predictor variables are linearly dependent on each other. This can lead to difficulty in analyzing the effect of variables on the target variable.

Explain the Anscombe's quartet in detail.

Anscombe's Quartet is the example to demonstrate the importance of data visualization and the importance of dealing with outliers, which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.

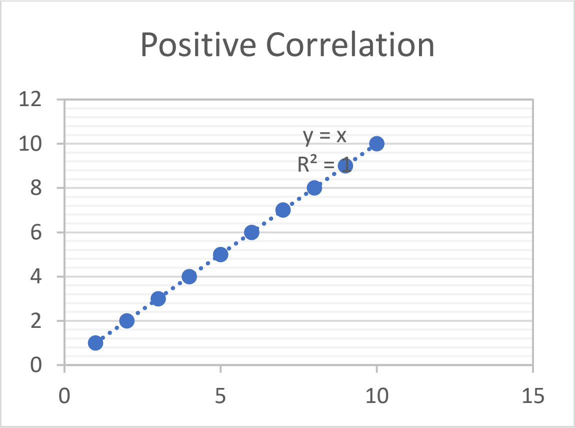
It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

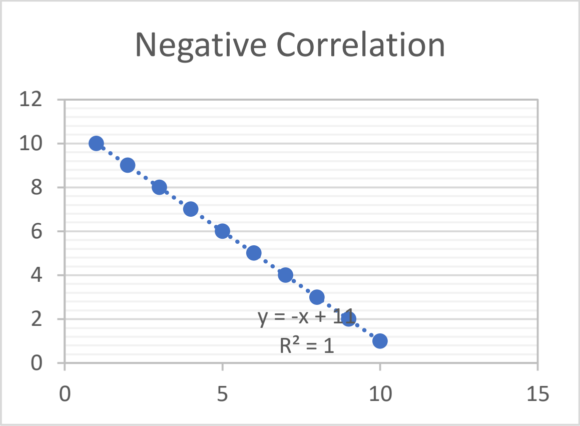


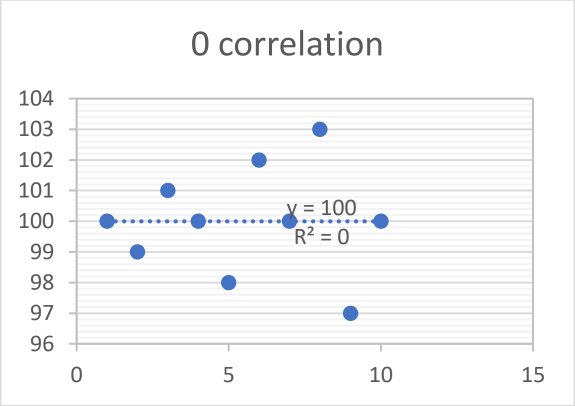
For the above 4 representations, the statistical parameters of mean ,standard deviation and correlation yet they are very different data sets. Anscombe's quartet shows the important of visualization.

What is Pearson's R

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. the Pearson's r is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1







Positive correlation implies that the two variables are positively related: one increases when other increases. Negative correlation implies two variables are negatively correlated , when one increases another reduces. 0 correlation is when two variables are not related to one another.

It is to be noted that Pearson’s r has two caveats: one is that it is applicable for linearly related variables and second is that correlation does not imply causation so even if one variable is highly correlated to another, it might be possible that both of them are dependent on a third variable.

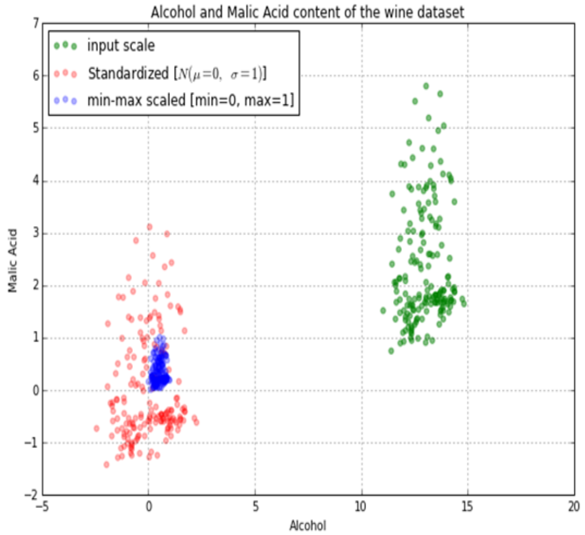
What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Feature Scaling also known as data normalization is a data preprocessing step. It is a method used to normalize the range of independent variables or features of data. Independent variables are of different types. We don't want our machine learning model to confuse a feature with a larger magnitude as a better one. Feature scaling in Machine Learning would help all the independent variables to be in the same range, for example- centered around a particular number (0) or in the range (0,1), depending on the scaling technique.

The first one is called standardized scaling and the second one is called normalized or min-max scaling.

Feature scaling is also important in case of gradient descent method, as it leads to a faster convergence. It also helps in making better interpretation of the model.

Normalized Scaling	Standardized Scaling
It is done by normalizing all the points on the data set to values between 0 to 1, by scaling it against the maximum and minimum values.	It is done by normalizing all data points to values against mean and standard deviation. It centers around mean and the standard deviation becomes 1 and mean becomes 0 after scaling.
It takes care of outliers since all the values are between 0 and 1.	It doesn't handle outliers since the values are not bounded.
It is preferred when data set does not follow Gaussian distribution.	It is preferred when the dataset follows a Guassian distribution.



The above diagram depicts the difference between standardized scaling and normalized scaling.

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

Variance Inflation factor (VIF) is the measure of multicolinearity in multiple linear regression.

$$VIF_i = \frac{1}{1 - Ri^2}$$

Where Ri^2 is the unadjusted coefficient for determining the regressing ith variable with respect to other predictor variables.

Value of VIF will be infinite when Ri^2 =1 , which is when ith variable is perfectly correlated with all the other predictor variables.

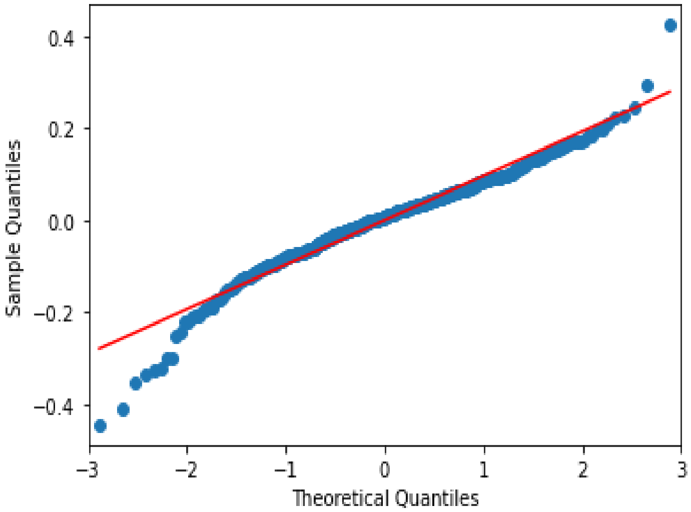
What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from same theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

It can be used in linear regression for 2 purposes:

i. During residual analysis, to see the fit of error distribution with normal distribution

ii. For confirming if test data and train data follow a similar distribution with the total population and hence can be a good representative for the population



The diagram shown above is the QQ plot for the residual of the Bike Lending assignment.