upGrad

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? (3 marks)

A. First listing down the categorical variables: ['season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit']

To analyze the categorical variables; I've used boxplot to understand the distribution with respect to the dependent variable.

Inferences observed of effects of categorical variables on the Target:

- Each year the booking has increased tremendously
- High sales were observed in Summer and Spring (S=2, 3) seasons
- Very low sales were observed when the weather condition is Rainy or Snowy
- Clear weather attracted more sales
- The months (April - Sept/Oct) had the high number of sales
- Sales were almost same on working or non-working days

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

A. Let's first understand what the argument i.e. drop_first=True means; while we are One Hot Encoding a categorical variable we use pd.get_dummies to create dummy variables, so this argument would basically drop the first dummy variable.

This helps in reducing the multi-collinearity as it would reduce the amount of collinearity between those dummy variables as if we have 4 levels; it can directly be represented with 3 dummy variables and the 4^{th} one would be a redundant variable; hence drop_first=True removes that extra variable i.e. return N-1 dummy variable (where N is the number of levels in the categorical column).

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

A. The numerical variable 'atemp' has the highest correlation with our target variable 'cnt' having a correlation value of 0.63.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A. The validations of assumptions of Linear Regression after building the model should be done as follows:

- Linear Relationship: There should be a linear relationship between independent variables and the target variable
- Normality of Error Terms: Error terms should be normally distributed with mean 0
- Multicollinearity: There should be no/insignificant relationship between independent variables/features
- Homoscedasticity: The errors terms should be random; i.e. there should be no pattern in the residuals
- Autocorrelation: No relationship between residuals

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

A. Following are the top 3 features which contributed significantly towards explaining the demand of the shared bikes:
   1. Temp
   2. Year
   3. WeatherSit3 (weather=3 (snowy & rainy))

   These features are determined as top 3 factoring in both the coefficients value from the model and the correlation to the target variable.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

A. Linear Regression is a regression algorithm in which we train and predict on continuous target variable. Some use-case scenarios for Linear Regression are: demand prediction, house prices prediction, etc.

   This algorithm basically builds a 2d line (for simple linear regression) where using independent variable (feature) we predict our dependent variable (target) by using a line equation.

   Mathematical Notation:

   $$Y = mX + c$$

   where,

   Y = target variable (output)
   X = feature variable (input)
   m = slope of the line (to be learned by algorithm)
   c = y-intercept (to be learned by algorithm)

   Here, using the given labelled data (X, Y), we build a best line equation using LeastSquares Method which minimizes the squared error; hence achieving a good regression model.

   Some assumptions that linear regression makes about the dataset are:
   - Linear Relationship: There should be a linear relationship between independent variables and the target variable
   - Normality of Error Terms: Error terms should be normally distributed with mean 0
   - Multicollinearity: There should be no/insignificant relationship between independent variables/features
   - Homoscedasticity: The errors terms should be random; i.e. there should be no pattern in the residuals
   - Autocorrelation: No relationship between residuals

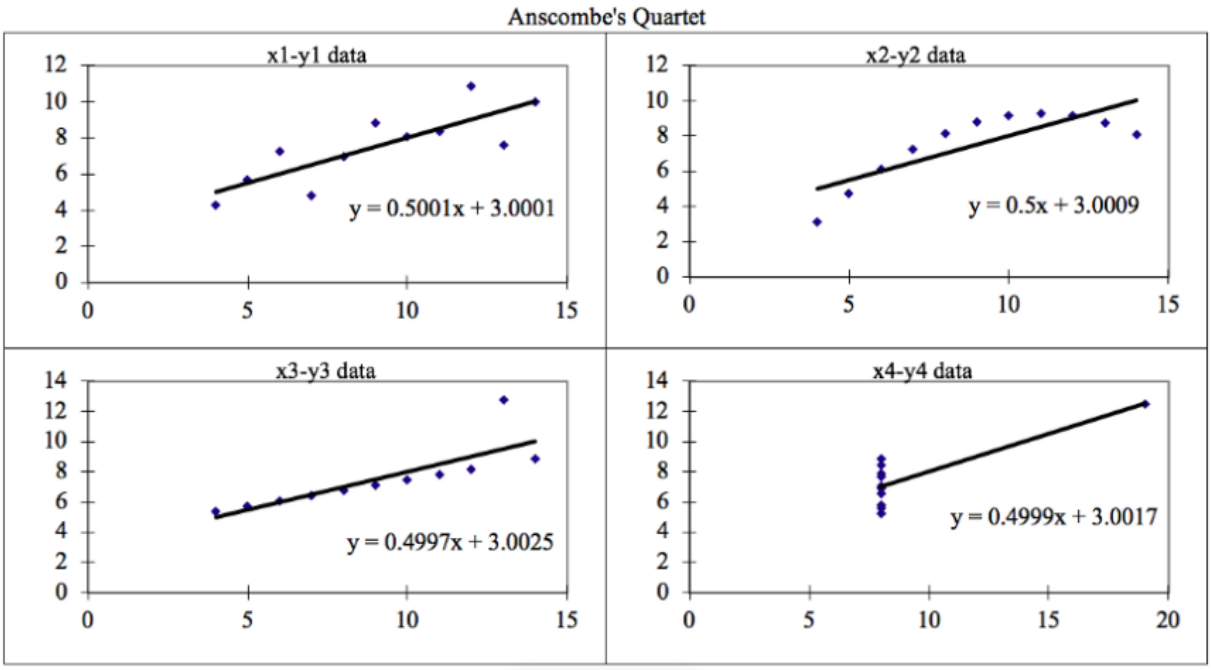
2. Explain the Anscombe's quartet in detail. (3 marks)

A. Anscombe's quartet consists of four data sets that have very identical simple descriptive statistics, but have very different distributions and appear very different when plotted.

   These four datasets are intentionally created to describe the importance of data visualization before passing the data to our Machine Learning model and shows how easy it is to fool our model.

Visualizing our dataset helps in:

- identify the various anomalies present in the data
- diversity of the data
- linear separability of the data
- relationships between variables



The four datasets can be described as:

- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

Application of Anscombe's quartet:

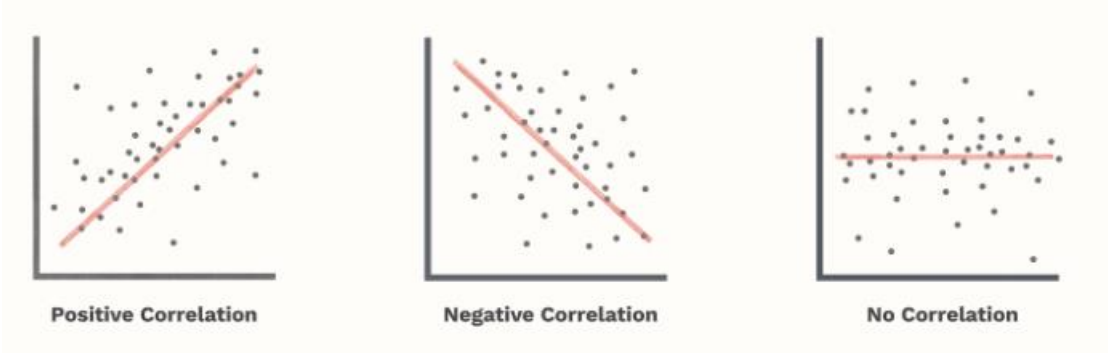
The quartet is used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R? (3 marks)

A. Pearson's R is a metric which is used to quantify the correlation between two variables. It is used to associate the strength attached to the linear association between the two variables.

This value ranges from [-1 to +1]:

- [-1 to 0] values: Negative correlation i.e. if our variable A goes up, then B goes down
- [0 to 1] values: Positive correlation i.e. if our variable A goes up, then B also goes up
- 0 value: No Correlation i.e. no relationship found between the two variables



The value shows the strength of the association between the variables i.e. 0.1 would mean that variables have a small positive correlation but whereas value of 0.9 would mean a heavy positive correlation and vice-versa.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

A. Scaling is technique where we scale the feature values present in the dataset into a fixed range.

This helps our linear regression model to not put heavy weights onto a particular feature just because its values are in a higher range.

This technique is performed before we feed our data into our Machine Learning model i.e. during Pre-Processing phase.

Normalized Scaling	Standardized Scaling
- This scaling uses the maximum and minimum values to scale the values	- This scaling uses the mean and standard deviation to scale the values
- The range of scaled values lies in [0, 1]	- Unbound range
- Used when we want all features values in a single range	- Used when we want 0 mean and constant standard deviation of 1
- Has the same distribution as the unscaled data i.e. if there were outliers in the old data; they'll still remain	- Isn't very much affected by outliers as we use the mean and std to scale the values instead of max & min values (as in normalized scaling) as in normalized scaling

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A. The reason for having VIF = Infinite is when we have a perfect correlation.

We have a high VIF when we have high correlation between the variables; so infinite basically means having the highest amount of correlation i.e. perfect correlation.

Mathematically understanding when a VIF = Infinite may occur:

$$VIF = 1 / (1 - R^2)$$

where,

R^2 = R Squared Value

So, whenever we get a R2=1, we get VIF = Infinite.
This is usually caused when we have some highly correlated variables in our dataset, so removing those variables would correct the VIF values.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A. Q-Q plot is an abbreviation for Quantile-Quantile plot.

This metric is a graphical technique to determine if two datasets come from populations with common distribution.

Usage:

- We use this plot to compare two dataset distributions to understand any similarities and dissimilarities they have
- We compare the quantiles of the both the datasets against each other
- Here quantile basically mean all the values below a given quantile value, example: if we have a quantile value of 0.25(25 percentile) this covers all the data below the 25^{th} percentile
- 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 $y = x$

Importance in Linear Regression:

- We can use the Q-Q plots to infer whether our training set and test set come from same population or not
- This helps to know the distribution changes in our test set (if there are any)

Interpretation of Q-Q Plot:

- Similar Distribution: If all points of quantiles lies on or near to straight line at an angle of 45 degree from x -axis