Name : Sreekanth VS

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)

   - weekday did not have much impact on the dependent variable

   - mnth had clear impacts on the dependent variable because month is highly correlated with the season

   - season has clear impact of the demand as well

   - weathersit has clear impacts on the demand

   - However, holiday and weekday do not seem to having as much impact as other variables.

   - Year 2019 has shown much more demand, this means the popularity is increasing steadily.

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

   - This helps to avoid redundant variables/features in the dataset

   - This also means one of the variable can be represented by all the remaining variables when creating dummies.

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

   - temp and atemp have highest correlation with the target variable.

   - However temp and atemp have a perfect linear relation amongst them. So one of them can be safely dropped.

- After this, windspeed has a moderate -ve correlation with the target variable.

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

   - Linear relationship between dependent and independent variables.
   - Perform residual analysis on the error terms.
   - All the assumptions regarding the error terms were validated
     - Error terms are normally distributed.
     - Error terms are independent of each other.
     - The error terms generally follow a normal distribution with mean = 0

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

     - Falls season, Winter season and Summer season has the biggest positive impact on the demands
     - Snow and rain has a strong negative impact on the demand

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   There are two types of Linear Regression, viz. Simple Linear Regression and Multiple Linear Regression.

   Simple Linear Regression is a model with only one independent variable.

   Multiple Linear Regression is a model with more than one independent variables.

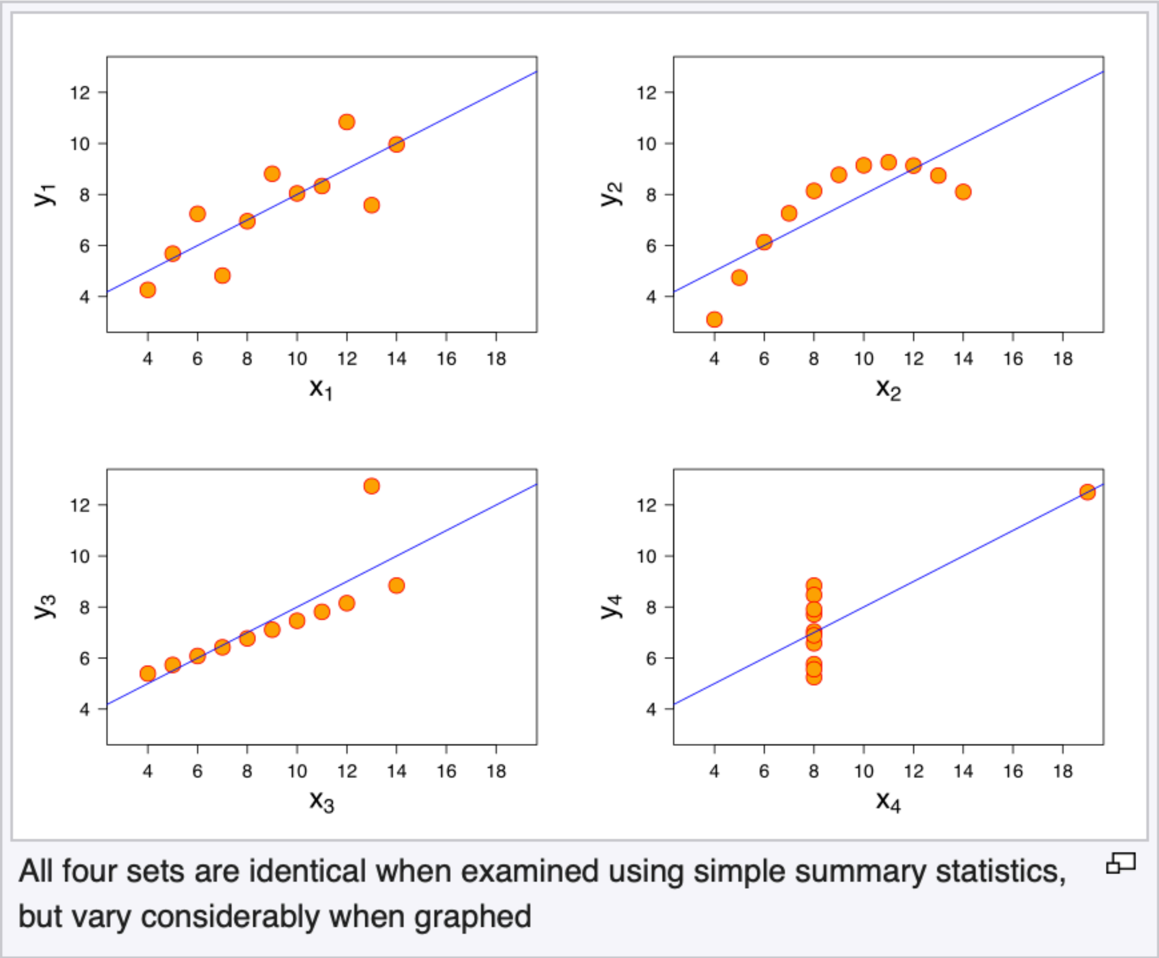
- Perform EDA on the dataset and get the data cleaned well
- Visualize data and analyze dependencies among various features.
- Convert categorical variables to dummies
- Perform train/test data split
- Scale continuous variables
- Perform outlier treatment
- Identify multicollinearity with respect to the independent variables.
- Reduce the number of features using automated RFE
- Build the linear regression model
- Check the statistics and further reduce the number of features using p-value
- Check the VIF and deal with all the features having VIF > 5
- Perform residual analysis for the error terms
- Predict target variable for the test data
- Check the r2 score for the test data and compare against train data.

2. Explain the Anscombe’s quartet in detail.

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.

- The first scatter plot appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

- The second graph; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph, the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



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

Source : Wiki

3. What is Pearson’s R?

In statistics, the Pearson correlation coefficient — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of 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.

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

   Scaling is the process of fitting data of continuous variables into a finite period. We need to scale features:

   - Ease of interpretation

   - Faster convergence for radiant descent model

   Scaling methods:

   - Standardization : The data is centered to zero with standard deviation of 1.

   - MinMax scaling : Get values between 1 and 0.

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

   If the r-squared value for a feature is perfect 1, then the VIF for that feature could be infinite.

   VIF = 1/(1 - r_squared)

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

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

In linear regression, if we have the training data set and test data set received separately then we can use the Q-Q plot to confirm that both the data sets are from the populations with the same distribution.