Bike Sharing Assignment

(Submitted By: Priyanka Tanpure)

Assignment – based Subjective Questions:

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

→ Bike Sharing booking increased and became popular in year 2019 than 2018 (from 'Year' variable) and were over 6000 bookings.

→ Bike Sharing Booking is more and preferred when the weather is clear (from 'Weathersit' variable).

→ Month preferred are May, June, July, August, September and October, almost nearly 5000 booking were done each month.

→ Fall and Summer seasons are more favourable for Bike Sharing Booking than Spring (from 'Season' variable) and booking were crossing 5000 bookings (median value).

→Weekday is showing more Bike Sharing booking almost between 4000-5000 bookings.

→ Workingday has more booking.

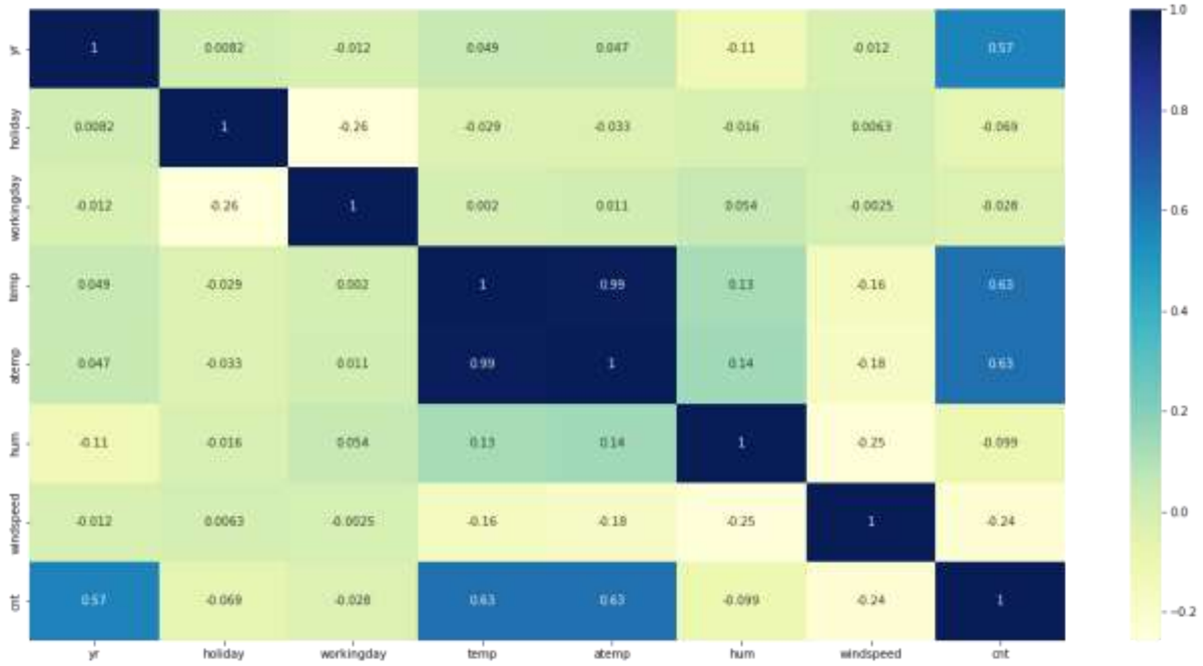
→holiday: non-holiday is more preferred for booking Bike Sharing.

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

→ To avoid Multicollinearity as it is undesirable and affect the model. By using drop_first = True as it will help in reducing the extra column created during dummy variable creation, hence it reduces the correlations among the dummy variables and also helps to avoid redundant features.

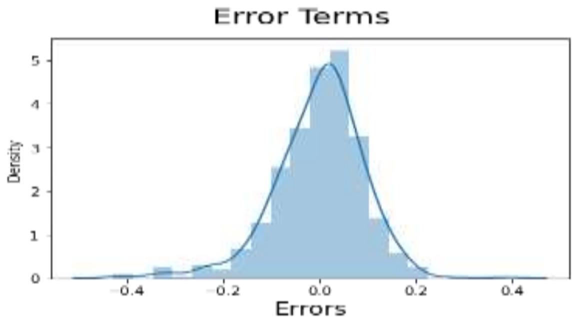
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

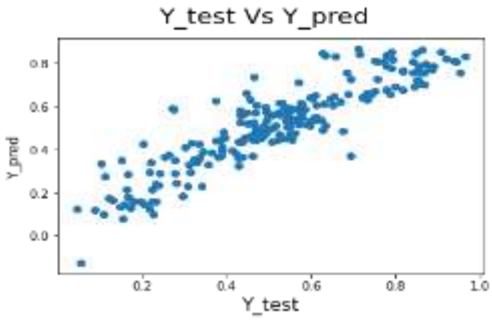
→ ‘cnt’ (Target Variable) has significantly high correlation with ‘temp’ (Temperature) = 0.63



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

→The assumptions of Linear Regression after building the model on the training set were validated by plotting the graph to see how the residual values(Error Terms) are distributed, they should be normally distributed which means model build was good.





OR we can simply plot a scatter graph between Y_test and Y_prep.

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

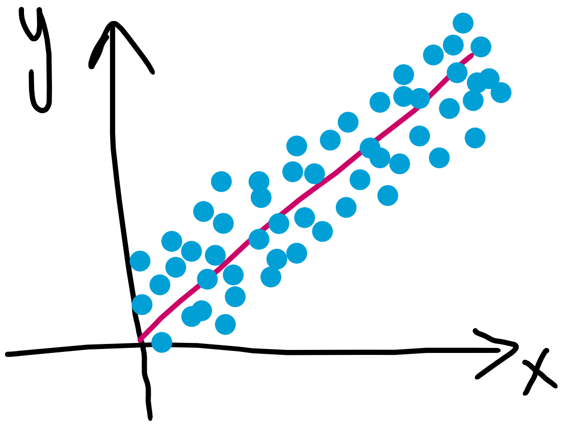
→ Top 3 features contributing significantly in the final model are:

- yr(2019) showing positive correlation
- temp showing positive correlation
- weathersit showing negative correlation

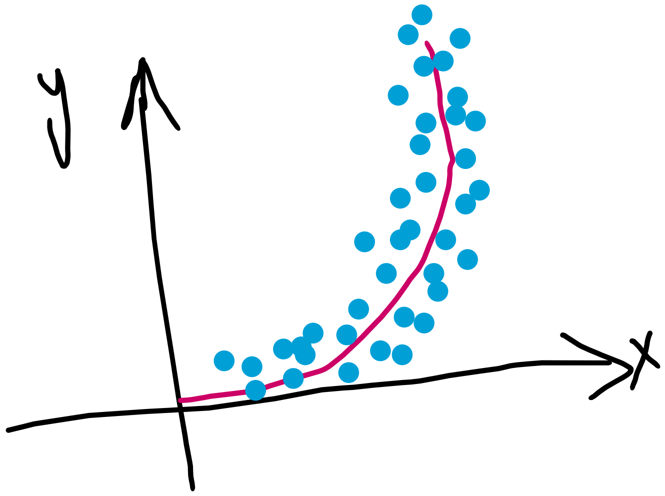
General Subjective Questions:

Q1. Explain the linear regression algorithm in detail.

→ Linear regression algorithm is a machine learning algorithm based on supervised learning category: it finds and creates the best linear-fit (or straight-line fit) relationship on given data between Independent (Target) and Dependent (Predictor) variables.



Linear Pattern



Non-linear Pattern

→ Depending on Independent variables the Linear regression models can be classified into two types:

1. Simple linear regression
   - Used when the number of independent variables is 1
   - Straight line is plotted using scatter plot to find the relation between Dependent and Independent variable.
   - The best line fit is $Y = \beta 0 + \beta 1X +\epsilon$
2. Multiple linear regression
   - Used when the number of independent variables is more than 1
   - The best fit line is $Y = \beta 0 + \beta 1X1 + \beta 2X2 +\ldots\ldots.+ \beta\rho X\rho$

→ The equation of the best fit line for linear regression is $Y = \beta 0 + \beta 1X$, found by minimising cost function. There are two methods used:

1. Differentiation

2. Gradient descent

→ The strength of a linear regression model is studied using $R^2 = (1 - RSS/TSS)$

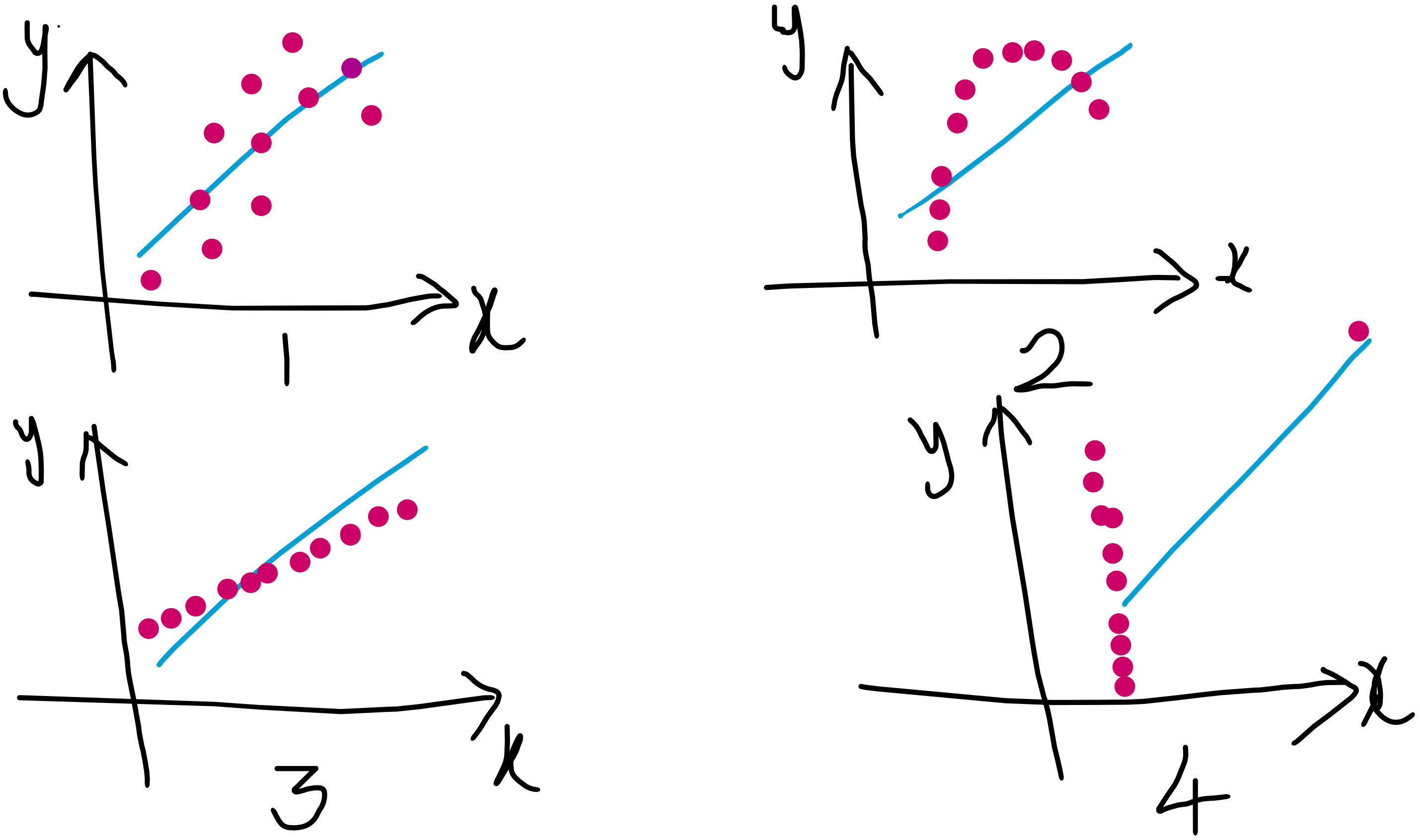
- RSS: Residual sum of squares
- TSS: Total sum of squares

→Mostly used method is Sum of squared residuals.

Q2. Explain the Anscombe's quartet in detail.

→ In 1973, Statistician Francis Anscombe explained the importance of plotting graph i.e. is visualization of given dataset before analysing it and also how statistical properties are affected due to the presence of outliers.

→ Anscombe's quartet has four datasets having similar statistical/descriptive properties, but when plotted in a graph they appear very different.



→ The dataset contains eleven (x, y) points each.

- Top left graph: Scatter plot shows linear relationship
- Top right graph: Non-linear relationship
- Bottom left: Perfect linear line for data
- Bottom right: One high leverage point to produce high correlation coefficient.

→ It tells the importance of data visualization before model building and helps to identify various outliers/anomalies present in the dataset like diversity, linear separability in data etc.

Q3. What is Pearson's R?

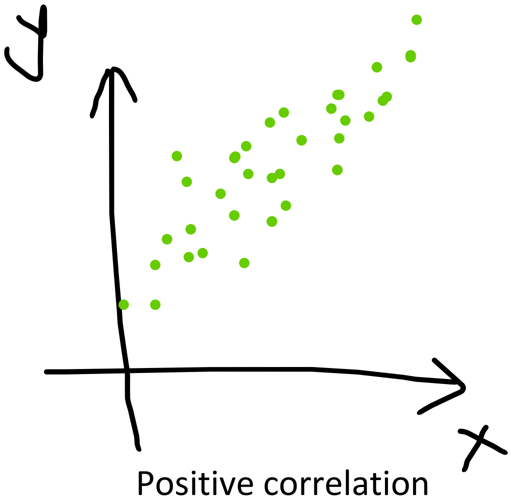
→ Pearson's R in statistics is also called Pearson correlation coefficient (PCC) or Pearson product-moment correlation coefficient (PPMCC) or the bivariate correlation: gives the linear relationship/correlation between given datasets.

→ Pearson's R = covariance of two variables/Product of their Standard deviation

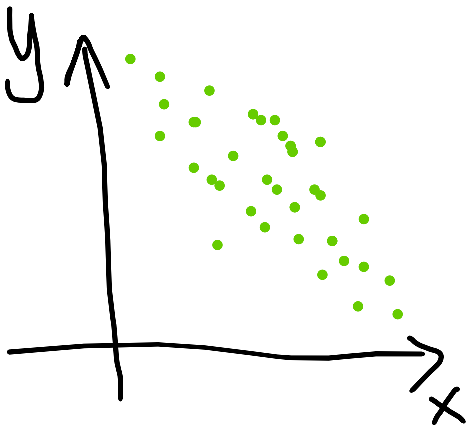
→The measurement of covariance is a normalized .

→The value of Pearson correlation coefficient is between -1 and +1:

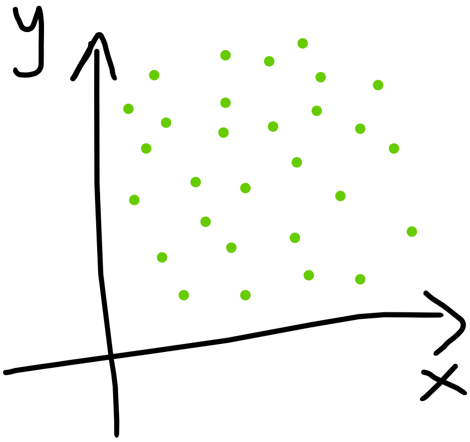
- r = 1: perfectly linear and positive slope
- r = -1: perfectly linear but negative slope
- r = 0: no linear relation/association
- 0 < r < 0.5: weak association
- 0.5 < r < 0.8: moderate association
- 0.8 <r < 1: strong association



Positive correlation



Negative correlation



No correlation

→ Formula: $r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}}$

- Correlation coefficient = r
- Sample values of x-variable = x_i
- Mean of x-variable = $\bar{x}$
- Sample values of y-variables = y_i
- Mean of y-variable = $\bar{y}$

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

→ Scaling is a step applied on data specially on the Independent variable so that the data can be normalized in a given particular range, so that the data is prepared for further use i.e. model building.

→It is generally performed in data pre-processing step.

→ Scaling is performed to make sure the models build are not biased in those high ranged features and are almost on the same scale and improve the performance of the model build in machine learning.

→Normalized and Standardized scaling:

S.No	Normalized Scaling	Standardized Scaling
1	Maximum and Minimum value is used	Mean and Standard deviation(SD) is used
2	Used when distribution is not known, features are of different scales and rescales the value in range [0,1].	Used when distribution is Normal or Gaussian i.e. Mean is zero and Standard deviation is one/unit variance.
3	Scale values between: [-1,1] or [0,1]	Not bounded by range
4	Outliers really affect scaling.	Outlier affect is much less.
5	MinMaxScaler transformer provided by Scikit-learn is used for Narmalization. i.e. sklearn.preprocessing.MinMaxScaler	StandardScaler transformer provided by Scikit-Learn is used for Standardization. i.e. sklearn.preprocessing.scale

	$x = \frac{x - min(x)}{max(x) - min(x)}$	$x = \frac{x - mean(x)}{s\,d(x)}$
6	Also called as Scaling Normalization.	Also called as Z-Score Normalization.

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

→VIF: is variance inflation factor and measures the collinearity among variables in multiple regression.

→VIF>10: high correlation/collinearity—remove/eliminate variable

VIF>5: moderate— worth inspecting

VIF<5: good value-- no eliminating/removing variable

→VIF = 1/(1-R_i^2) where 'i' refers to the ith variable

→When the VIF is infinite it represents perfect correlation/perfect multicollinearity i.e. the association between the two independent variables is highly collinear i.e. the value of R is equal to 1 so it makes VIF infinity as the denominator becomes zero.

- It happens when a corresponding variable and other variable are expressed exactly in a linear combination.

→To solve this we need to drop the variable causing the perfect multicollinearity.

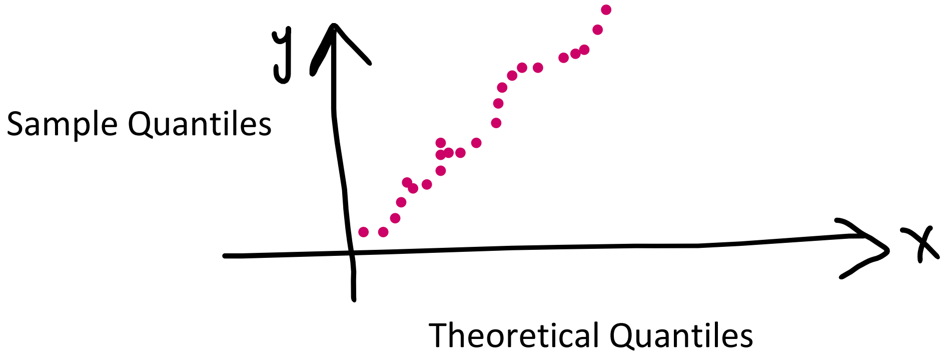
Q6. What is Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

→Q-Q plots: stands for Quantile-Quantile plots which is a graphical method used for plotting two quantiles against each other.

- It is a scatter plot.
- Quantile has some values above it and some below.

→Quantiles of a sample distribution (SQ) and a theoretical distribution (TQ) are plotted against each other to determine if the two datas belong to same/common distribution.

→If the quantiles belong to the same distribution the line/plot/graph formed is roughly a straight line similar to plot obtained from a normal distribution.



→Uses of Q-Q plot:

- Using this its possible to identify the distribution types
- Shapes of the distribution can be compared using Q-Q plot also it can determine if two population/data are of the same distribution
- Can determine the properties like location, scale skewness of distribution(similar/different)

→Importance of Q-Q plot:

- While working on two samples (similar or different) the sample size need not be same/equal.
- Simultaneously many different distribution properties can be determined like: location, shift in scale, outliers presence, symmetry changes etc