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. Following are the categorical variables and their effect on dependent variable cnt in the dataset, which we analyzed using Boxplot:

- ✓ Season - Fall have the highest booking, and spring have the lowest booking. Summer, Fall and Winder are almost closer. That means in Spring bookings are least
- ✓ Holiday - rentals reduced during holiday.
- ✓ Mnth - Booking increase slowly from Jan to July and then stay highest till Sept and then slowly starts decreasing (it might be related to seasons) September had highest number of rentals while December had least. This observation is on par with the observation made in weathersit. The weather situation in december is usually heavy snow.
- ✓ weekday: Weekday mostly not affecting bookings
- ✓ workingday: Working daty it does not have much effect
- ✓ weathersit - Weather situation heavily impact the bookings, if weather is clear bookings are more and based on weather conditions it decreases
- ✓ Yr - There is noticable increase in bookings from 2018 to 2019

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

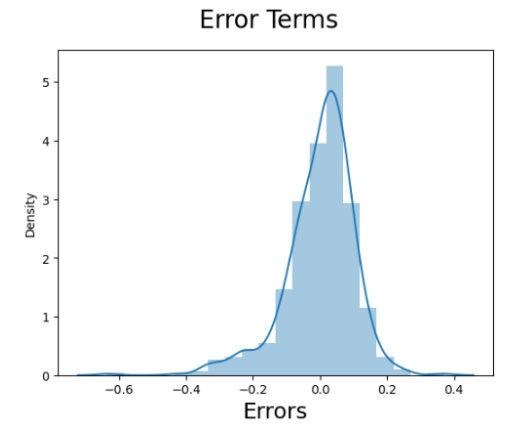
Ans:. All categorical features can be represented as N-1 dummy variables, where N is the number of categories. If we don't drop the first column, then our dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may have trouble converging and lists of variable importances may be distorted. Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we lose one colum

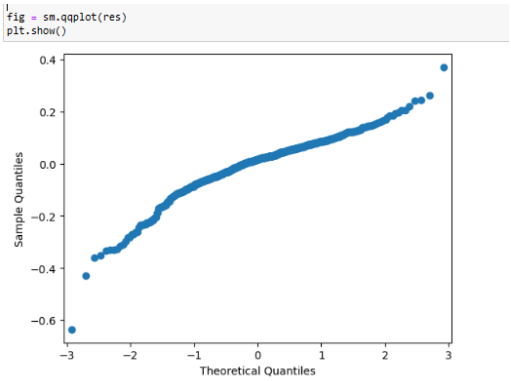
3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

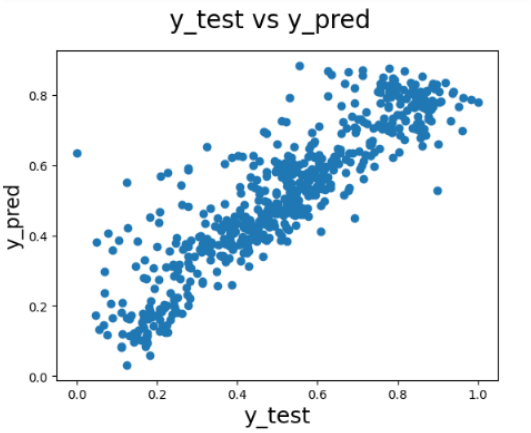
Ans: From pair plot we can see temp and atemp have highest correlation with the target variable. Where temp and atemp are correlated to each other and add multicollinearity.

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

Ans: Below are the evaluation done to confirm assumptions of Linear Regression are followed:







We plotted Residual as histogram and Q-Q plot to confirm if that the model is following Linear Regression assumptions.

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

Ans: Based on the final equation we can see demand of bike is mostly explained by temp(`0.433354`), yr (`0.241186`) and windspeed(`-0.135404`).

General Subjective Questions

1: Explain the linear regression algorithm in detail.

Ans: Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x).
Linear Regression is of two types: Simple and Multiple.
Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

Equation of Simple Linear Regression, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

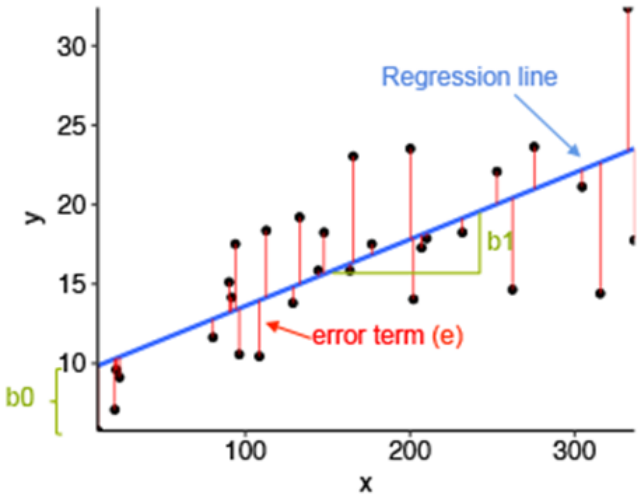
Equation of Multiple Linear Regression, where bo is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Let's understand this with the help of a diagram.



$$y = b_o + b_1 x$$

Image Source: Statistical tools for high-throughput data analysis

In the above diagram,

- x is our independent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- b_o is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.
- The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

Mathematical Approach:

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2

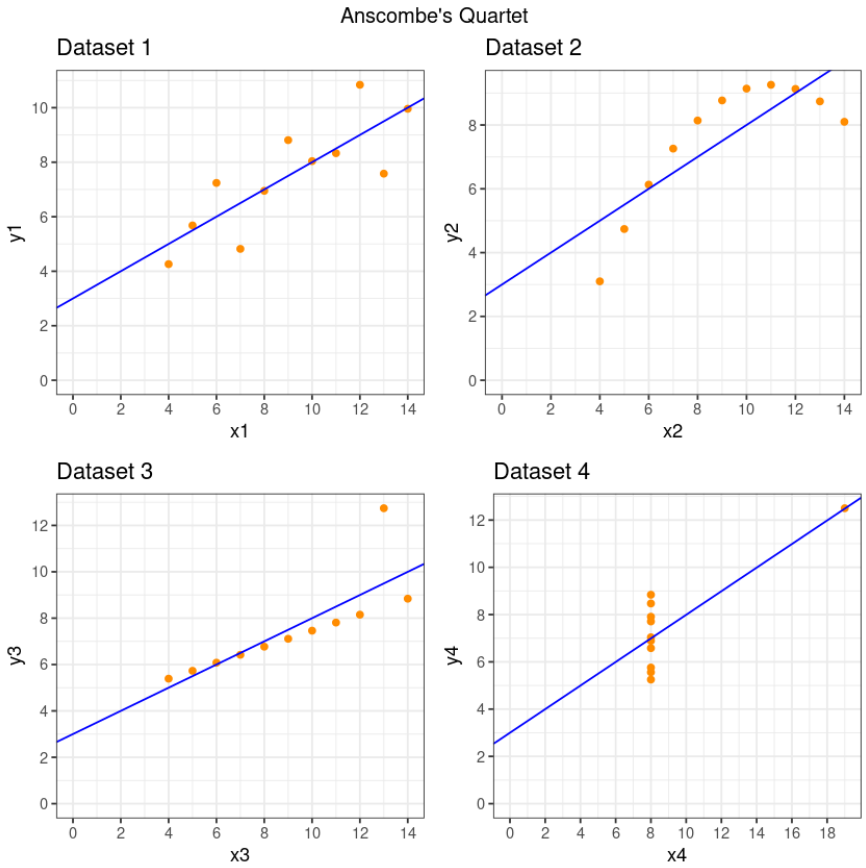
i.e

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

Rsq, AdjRsq, MSE,RMSE,MAE – 5 evaluation metrics

2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them,it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line 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 (bottom right) 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.

3: What is Pearson's R?

Ans: Pearson's R is a numerical summary of the strength of the linear association between the variables. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data.

- ✓ R = 1 means the data is perfectly linear with a positive slope
- ✓ R = -1 means the data is perfectly linear with a negative slope
- ✓ R = 0 means there is no linear association

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

Ans: *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.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

Ans: VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity.Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

Ans: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.It is used to compare the shapes of distributions.A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

- ✓ Do two data sets come from populations with a common distribution?
- ✓ Do two data sets have common location and scale?
- ✓ Do two data sets have similar distributional shapes?
- ✓ Do two data sets have similar tail behavior