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? (3 marks)

- When the weather is clear there are more bike rentals, and very less during the rainy season.
- Fall seems to be the most preferred season for bike rentals followed by summer. Spring season has least bike rentals.
- We can say that clear sky during fall/summertime will be the best time for rentals.
- Light Snow, Light Rain +Thunderstorm + Scattered clouds, Light Rain + Scattered clouds during Spring/Summer is the worst time for rentals.

Q2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

If we have variables with k different values, then we actually need only k-1 dummy variables to show all the values of k variables. So we drop one variable using drop_first=True.

If we don't drop it, then one variable value can always be predicted based on the rest k-1 variables and this will lead to multicollinearity. This will impact our interpretation of the model.

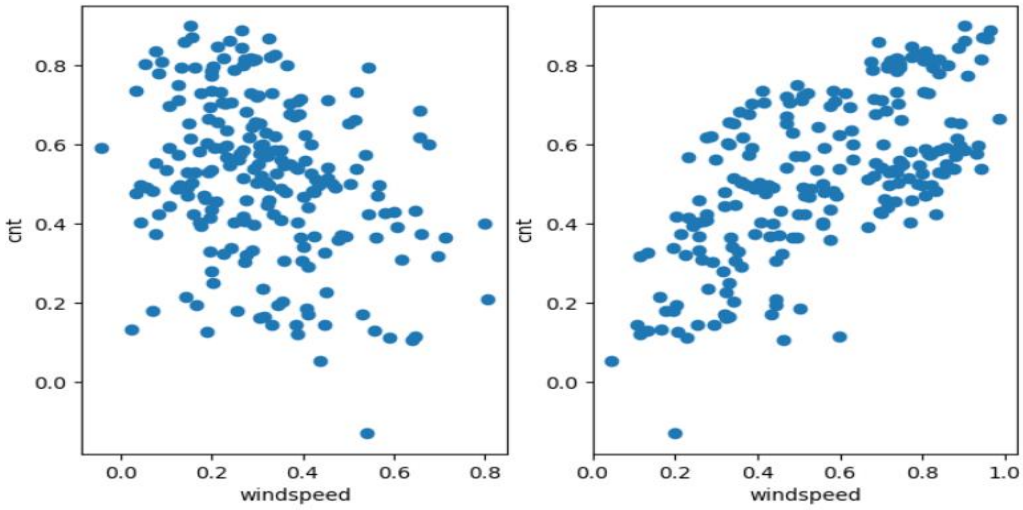
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Its registered users.

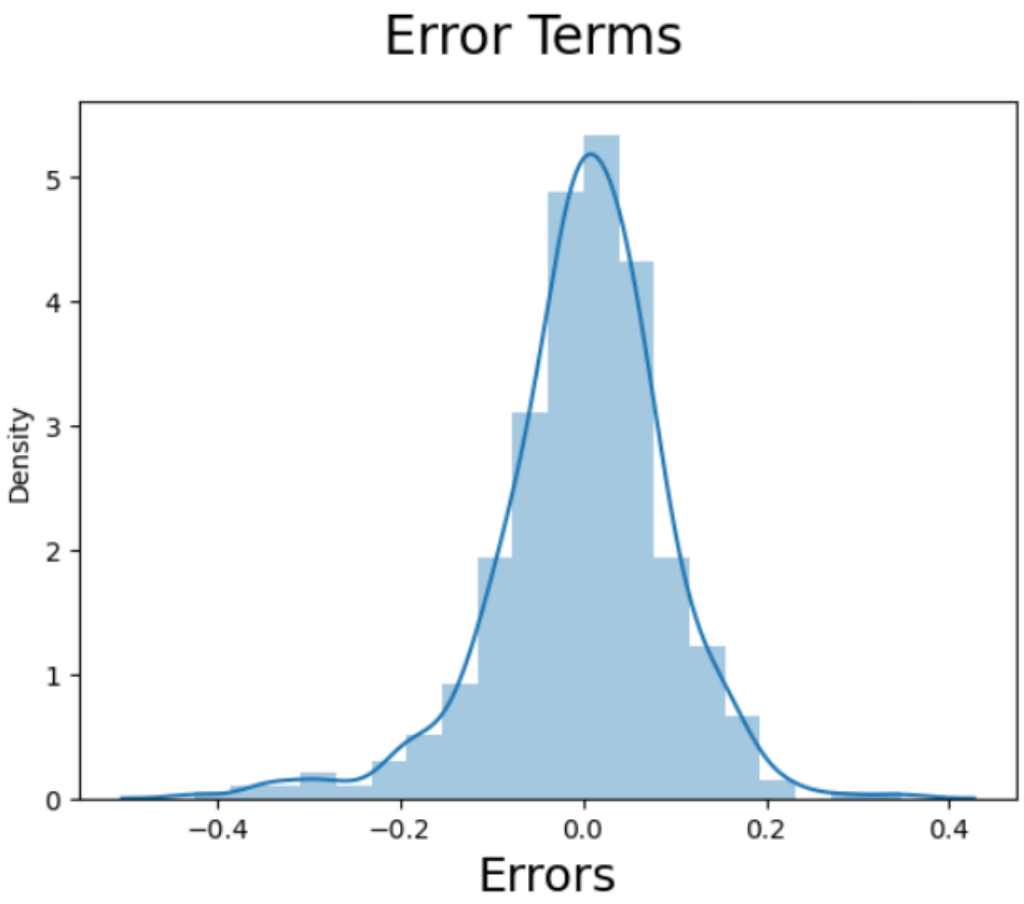
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- The relation between the target and predictor variables in linear. We saw from the model equation that it's a linear model. Scatter plot shows the linearity of the variables.

  Variables with least and most coefficient values



- The distribution of error terms is normal and centered around 0, that confirms the Homoscedasticity assumption.



- Multicollinearity was taken care by removing columns with more than 5 VIF.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

I. Temperature is the most significant feature according to the mode. It has a coefficient of 0.4669.
II. Year is another significant variable. There has been an increase in rentals from 2018 to 2019.
III. Rain (or Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) is negatively corelated and is one of the significant variables for rentals demand.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression algorithm is basically an equation for a straight line –

y = B0 + B1X1 + B2X2 +…

- Here y is the dependent variable.
- B0 is the intercept or the constant, which is the value of y when rest of the predictor variables (X1, X2 …) are 0
- B1, B2 is the slope of the line, which signifies how strong is the relation between predictor variable and dependent variable. B1 means, y will increase B1 times for each unit increase in X1 given rest of the values are constant. Similarly, for B2, B3 etc.

The aim of the algorithm is to create a straight line which fits best into our data. The line then will be used to predict the values of y (dependent variable) based on new data/ predictor variables.

Best Fit is determined by the least residual (difference between the actual y value and the predicted value of y) or it's the least error. We use that value of B1, B2 which minimizes the Residual Sum of Squares (sum of squares of different residuals)

Our Final Algorithm based on the model is –

cnt = 0.1813 + 0.2351*yr -0.0638*holiday + 0.0523*weekday + 0.0190*workingday + 0.4669*temp - 0.1552*windspeed -0.0775*Mist -0.2828*Rain -0.0813*spring + 0.0402*summer +0.0782*winter

Here B0 is 0.1813. Different predictor variables are year (yr), holiday, weekday etc. and their coefficients/slope are along with them.

Q2. Explain the Anscombe's quartet in detail. (3 marks)

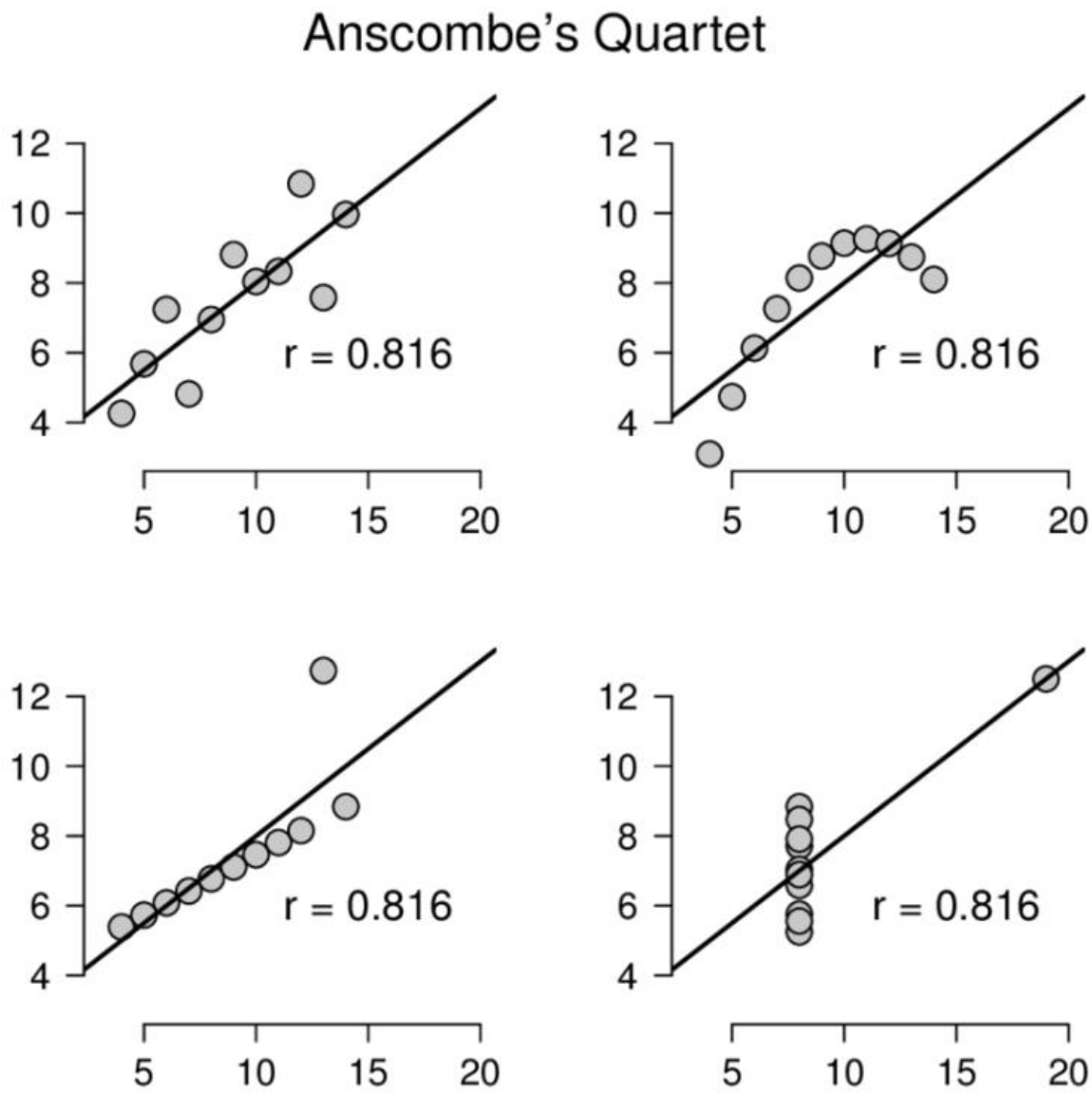
Anscombe's quartet is a combination of 4 datasets (values of x , y) which show that statistics do lie ☺

The datasets have almost same values for different statistical calculations like –

- Mean of x and y
- Variance of x and y
- Correlation between x and y
- R square value
- And the linear regression question.

By looking at these values one would think that all the datasets are similar and one can be replaced with other for any analysis. Or if these datasets belonged to a business from 4 different locations and we were to check which one is performing better, based on statistics we could infer that all the locations are performing in the same way.

But when the datasets are plotted (scatter) over a graph, its clearly visible that the 4 datasets are very different.



Link to - SourceforImage

i. Top Left is the simple linear relation.
ii. Top Right is a nonlinear relation, the curvature changes the correlation here.
iii. Bottom left is again a linear relation with an outlier. Here the outlier correlation the stats.
iv. Bottom right shows that the even though there is no relation between x and y except for one variable. But that variable values are good enough to change the correlation of the dataset.

So its important to visualize the data too.

Q3. What is Pearson's R? (3 marks)

It is the correlation coefficient that is used to measure the correlation between two variables.

The value for Pearson's R (r) can lie between -1 and 1.

- A value between 0 to 1 shows a positive correlation. That means as one variable increases another variable will also increase and if one decreases so will other.
- 0 value means there is no correlation, the variables are independent of each other.
- 0 to -1 means a negative correlation, so if a variable increases other decreases and vice versa.

An absolute value greater than 0.5 is considered a Strong correlation.

0 to 0.3 is a weak correlation. Otherwise, a moderate one (*other than 0 where we don't have a correlation*)

Formula to calculate Pearson's R is –

r =((sum of x * y) – (sum of x) * (sum of y))/ square root of (((sum of square of x) – (square of sum of x)) * ((sum of square of y) – (square of sum of y)))

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

Scaling is used to make the variables available for comparison and to be used together in a linear regression model. Variables with different scales are changed to come to same scale, either between -1 and 1 (in case of Standardized scaling) or between 0 and 1 (in case of Normalized scaling)

Without scaling the interpretation of the model becomes difficult. The coefficient of the variables without scaling will not give a true picture. A variable on a large scale will have smaller coefficient and vice versa. And based on coefficient we can wrongly interpret the relation between the variables.

Also, the optimization of the model also becomes faster by scaling.

- Normalized scaling, also called Min-Max Scaling. It compresses the variable data between 0 and 1.It's calculated by (x – x_min)/(x_max-x_min), where x is the variable and x_min is the minimum value of x and x_max is the maximum value of x. This will limit all data between 0 and 1.
- Standardization converts data between -1 and 1. Its calculated by (x – x_mean)/sigma. This will limit data between -1 and 1.

Eg: one variable windspeed has values from 10 to 1000 and other variable temperature has values from -5 to 40. By scaling we can compress all variables between 0 to 1 or -1 to 1. This will help in placing them in one linear equation and correctly interpreting their coefficients.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Infinite value of VIF means that two variables have a perfect correlation. The R square in such case will be 1. And because of VIF = 1/ (1 – R square), it comes out to be infinite.

This can happen in case of duplicate values. Eg – Sum of casual and registered bikers and Cnt of bikers (which actually is a sum of the two bikers)

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

Q-Q plot is a Quantile-Quantile plot. Quantiles are cut points dividing the range of probability distribution into continuous intervals with equal probabilities. One dividing into 2 equal parts is called a median and into 100 equal parts is percentile.

Q-Q plot is created by plotting scatter plot from 2 data quantiles against each other. If we know the other distribution type (normal, uniform etc..), the Q-Q plot will tell us which distribution does our data matches to.

We can create a Q-Q plot between the quantiles of 2 datasets and see how similar the distributions are. So it helps in comparing to distributions. This is helpful in machine learning to see if train test distributions are same.