

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)

Answer: The categorical variable in the dataset were season, mnth, weathersit, holiday, weekday and weekday. These were visualized using a boxplot. These variables had the following effect on our dependent variable:-

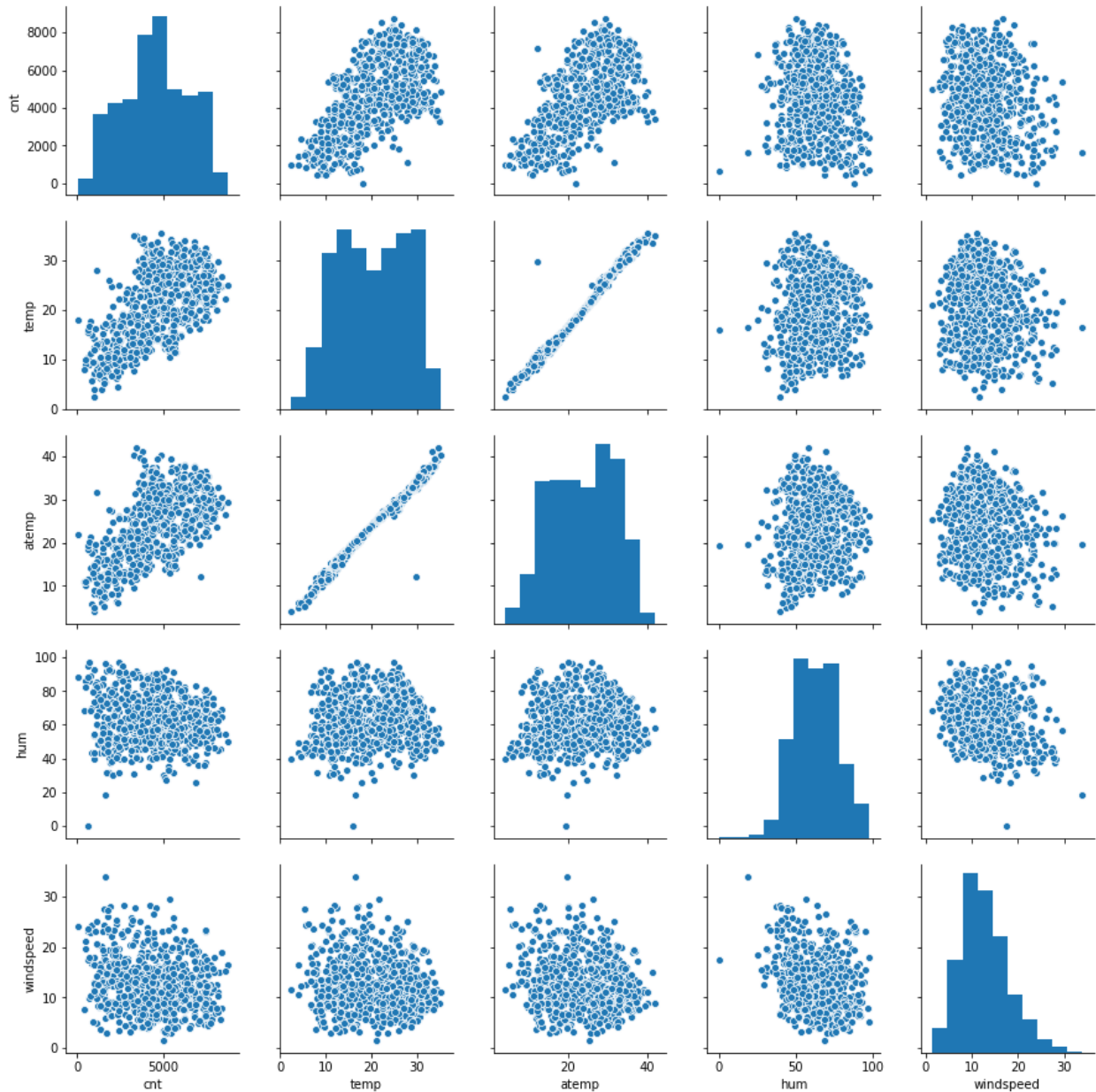
- 1. Season:** Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
- 2. Mnth:** Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
- 3. Weathersit:** Almost 67% of the bike booking were happening during 'weathersit1' with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
- 4. Holiday:** Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- 5. Weekday:** weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
- 6. Workingday:** Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

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

Answer: If you don't drop the first column then your 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 column.

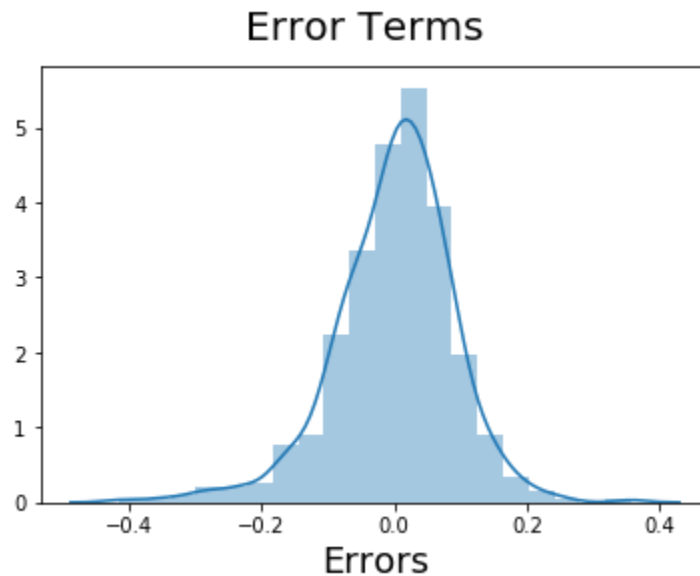
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Using the pair plot, we could see there is a Linear Relation between temp and atemp variable with the predictor 'cnt'.



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

Answer: Residuals distribution should follow normal distribution and centered around 0. (Mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.



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

Answer: The top 3 features are

1. Temp - coefficient: 0.491508
2. yr - coefficient: 0.233482
3. Weathersit_Light Snow & Rain - coefficient -0.285155

General Subjective Questions

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

Answer: Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation “ $y = mx + c$ ”.

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

- 1. Simple Linear Regression: SLR** is used when the dependent variable is predicted using only **one** independent variable.
- 2. Multiple Linear Regression: MLR** is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots$$

β_1 = coefficient for X1 variable

β_2 = coefficient for X2 variable

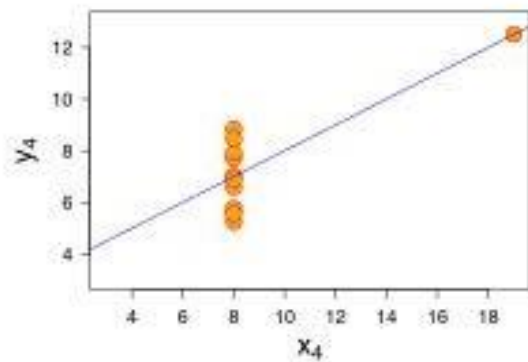
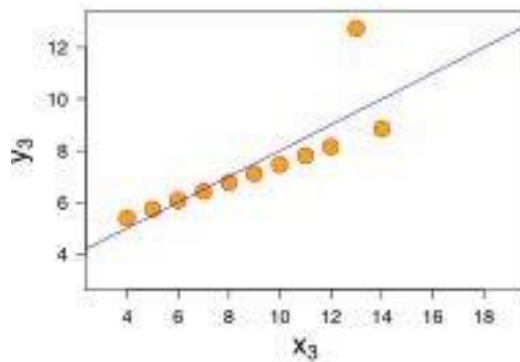
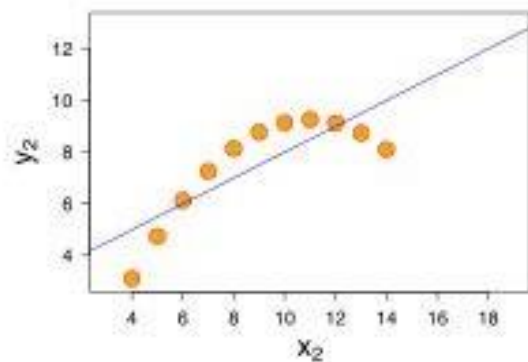
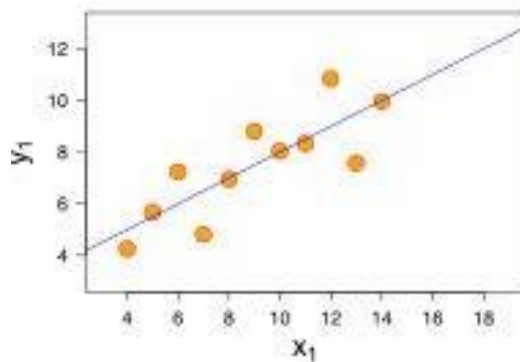
β_3 = coefficient for X3 variable and so on...

β_0 is the intercept (constant term).

2. Explain Anscombe's quartet in detail.

Answer: 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? (3 marks)

Answer: 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? (3 marks)

Answer: Feature **scaling** is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Answer: **VIF - the variance inflation factor** - The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. $(VIF) = 1 / (1 - R^2)$. If there is perfect correlation, then $VIF = \text{infinity}$. Where R^2 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 its 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. (3 marks)

Answer: 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?