Subjective questions

General Subjective Questions

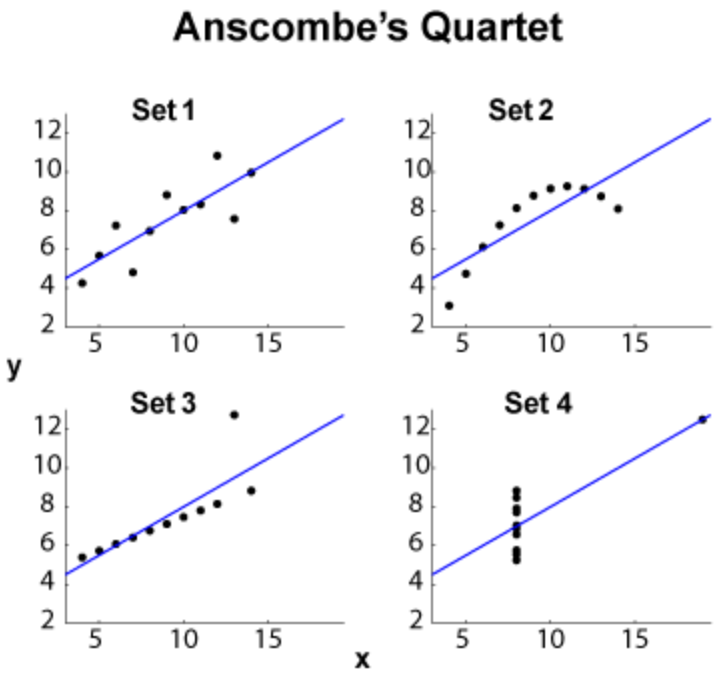
1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm which performs regression task. A regression model predicts a continuous value, i.e the target variable is a continuous variable. Ex: Prediction of housing prices, prediction of sales for the marketing budget,etc. Linear regression predicts a dependent variable y based on one/many independent variables x_i.It basically fits a line/ hyperplane model depending on the independent variables to predict the value of y. A simple linear regression model is a line equation, $y = mx + c$, where c is the intercept and m is the slope or coefficient.The main aim of the model is to get the optimal value of m , such that the line obtained based on the equation predicts the value of y for a new x with minimum error. Optimal value of m/coefficient can be obtained by minimizing the cost function. There are many cost functions, like root squared error, ordinal least square error etc. Basically a cost function computes the error, and help us to reach the optimal solution.There are two most common ways to minimize the cost function:

- Using differentiation
  If $J(\Theta)$ is the cost function, then $\frac{\delta(J(\Theta))}{\delta\theta}$=0 will get the optimal value.
- Using gradient descent algorithm.
  Gradient descent is an optimization algorithm, which descends the slope in each iteration to reach the minima. If α is the learning rate , then in each iteration new value of x is calculated based on α and $J'(\Theta)$.This process is continued until the value of x no longer changes / changes insignificantly with next iteration.

2.Explain the Anscombe's quartet in detail.

Statistics like mean,variance,mode can help describe the data. They help us to understand how much variation is there in the data, without actually looking into all the data points. But statistics alone can't fully depict any dataset. Francis Anscombe on realizing this created several datasets with identical statistical properties to demonstrate this. These datasets and graphs are called Anscombe's quartet.



The figure below shows the graphical representation of the four datasets. It has to be noted that even though they have the same variance,mean and linear regression, the nature of the datasets are different. It can be seen here even though the four datasets have the same linear regression, the top left graph shouldn't be analyses with linear regression due it's curvature. Anscombe's quartet basically emphases on graphing the data before the analysis to understand the nature of the data better.

3. **What is Pearson's R?**

Correlation helps us determine the strength of linear relationship between two variables. There are two types of correlation: parametric and non-parametric. Pearson's correlation is a parametric correlation, which gives the strength of linear dependency between two numeric variables.The formula for calculating pearson's coefficient is given below.

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2] - [N\Sigma y^2 - (\Sigma y)^2]}}$$

The value of r lies between -1 to +1.

- Negative correlation means - if one variable increases, the other variable decreases.

- Positive correlation means: if value of one variable increases, then the value of other variable also increases.
- If the value of r is closer to 1 or -1, the two variables are strongly related with each other.

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

Scaling is a technique to transform the data into a common range of values. Scaling helps to speed up the gradient descent process. Gradient descent is an optimisation algorithm which helps to find the best fitted model with minimum error between the predicted and actual value. In each iteration the algorithm descends in the direction until it reaches minima.The slope(Θ) will descend quickly on smaller ranges rather than large ranges. In case of large ranges, the algorithm will oscillate inefficiently, and might take some considerable amount of time to reach optimum value.

There are two common techniques in feature scaling :

- Normalized feature scaling
- Standardized feature scaling

Standardized	Normalized
$x = x - \frac{mean(x)}{SD(X)}$	$x = \frac{x - min(x)}{max(x) - min(x)}$
There is no restriction on the range of x.	x is in range of [0,1]
Outliers can still be identified even after scaling	Difficult to identify outliers , as all the data points are in the range of [0,1]

Standardized data will have a mean of 0 and SD of 1.

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

Variance inflation factor is used to check if a feature variable has multicollinearity with other feature variables. The intuition behind this is to build a model to predict x_i using other predictor variables. VIF is calculated using the formula below.

$VIF_i = \frac{1}{1-R_i^2}$

If value of VIF is high, then it means that variable x_i is highly correlated with other variables. **If value of VIF is infinity, then R_i^2= 1, which means that there is a perfect correlation with other feature variables.**

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

Q-Q plot(Quantile-Quantile plot) is a graphical tool which helps to determine of two datasets come from the same distribution such as normal,exponential,etc. It is a probability plot of quantiles of the first dataset against the quantiles of the second dataset. Interpretation of q-q plot is as follows:

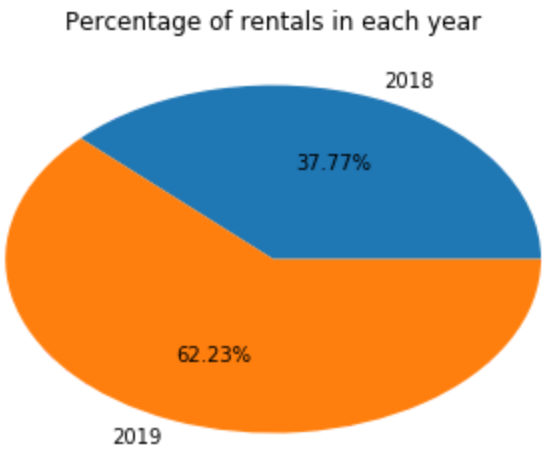
1. 45 degree line - If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
2. Above or below 45 degree line - If the two data sets have come from populations with different distributions then the data points will be far from the reference line.

The Q-Q plot is used to see if the two datasets come from the same distribution. This is useful in case of linear regression, as we have two datasets - train data, and test data. Linear regression can help to predict /forecast values only if they are in the same range as of training data. Values outside the range cannot be predicted using linear regression. So q-q plot plays an important role here to check if both the training and test data come from the same normal distribution.

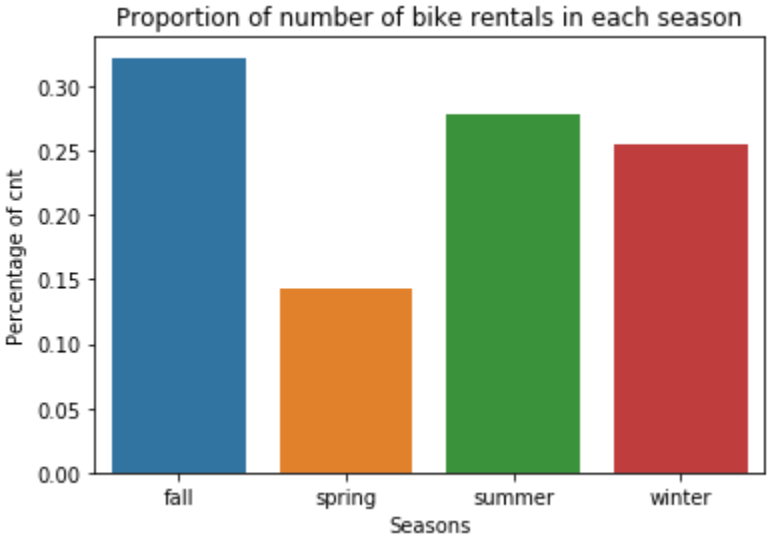
Assignment-based Subjective Questions

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

- Year column - from the analysis the year 2019 saw higher bike rentals than the year 2018.



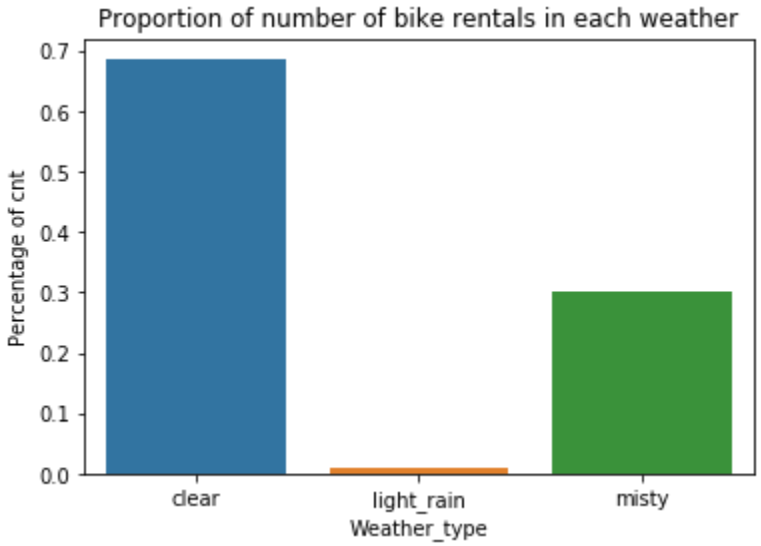
- Season - There are four seasons namely fall,summer,spring and winter. Fall has the highest proportion of total bike rentals.



- Weather

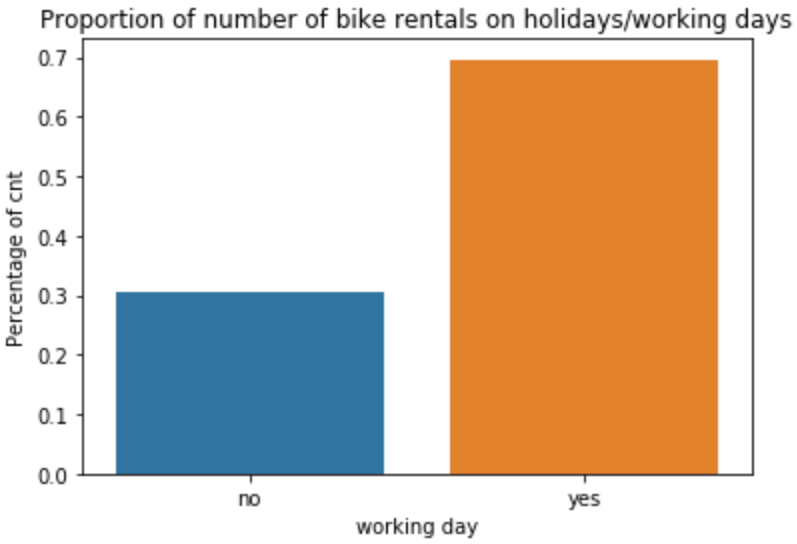
Actual value	Interpretation	Mapped value
1	Clear, Few clouds, Partly cloudy, Partly cloudy	clear
2	Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist	misty

3	Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds	Light rain
4	Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog	heavy



- Clear weather has a higher percentage of bike rentals.

- Months - august,june,july and september have the highest bike rentals.
- Working day - if day is neither weekend nor holiday is 1, otherwise is 0. The bike rentals are higher on working days, mostly they are used by people to commute to work.



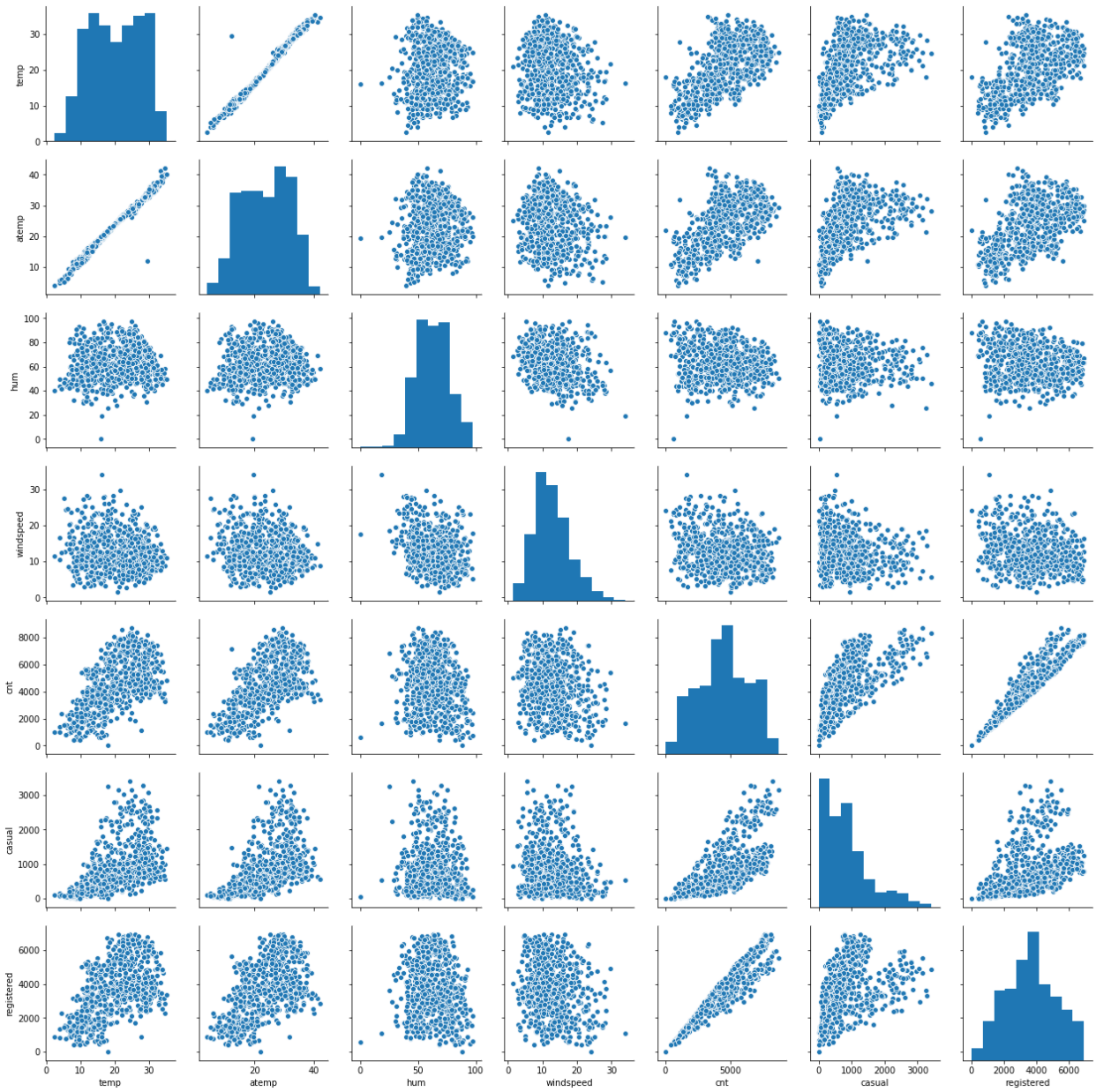
2. Why is it important to use drop_first=True during dummy variable creation?

For n distinct values, we need to create n-1 dummy variables to represent the data.If we don't pass drop_first=True param, then it will create n dummy variables. There will be one extra variable which is redundant. For ex; in the bike sharing system dataset, season has column has four values namely,fall,summer,spring and winter. If we pass drop_first=True, there will be three columns: summer,spring and winter. It is fall when summer,spring and winter all have 0 value.

spring	summer	winter
0	0	0

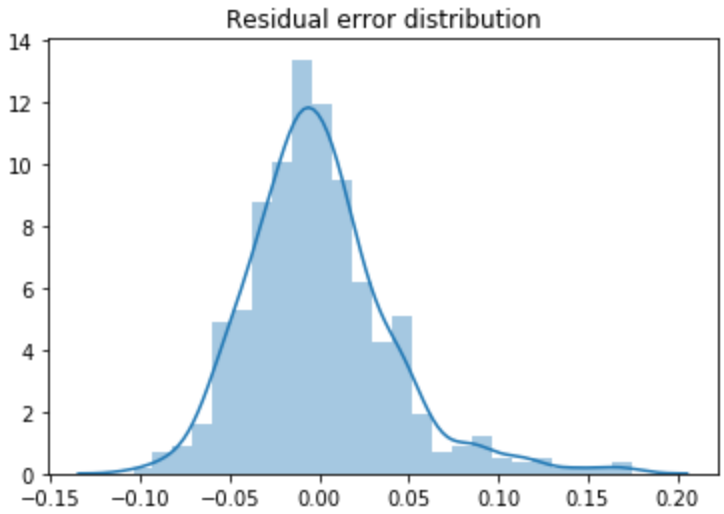
3.. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The feature variables temp, and atemp have the highest correlation with the target variable cnt. Cnt also has correlation with registered users.



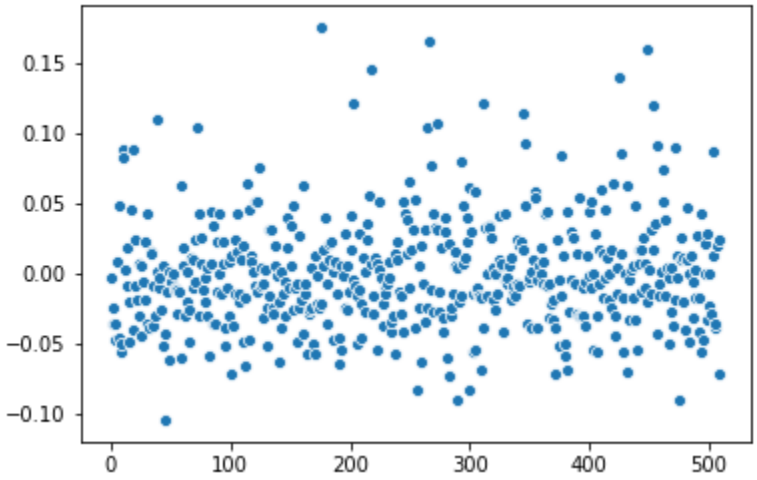
4 How did you validate the assumptions of Linear Regression after building the model on the training set?

1. Distribution plot of error terms
   According to the assumptions of linear regression, the error terms should have a normal distribution with mean 0. The figure below validates this assumption.



2. Scatter plot of error terms

According to the assumption the error terms should not be any dependency among the error terms, i.e there should not be any pattern found in the error terms.



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

- Temp
- Year
- Light rain