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. After building a model with predicted R-squared of 0.80 the model had following equation:
cnt = 1963.58 + 4071.06 * temp - 1321.15 * windspeed - 709.44 * spring + 666.435 * winter + 338.21 * summer - 654.24 * Mist+Cloudy + 2038.50 * yr - 741.15 * holiday - 2446.46 * Snow+Rain - 239.20 * tuesday

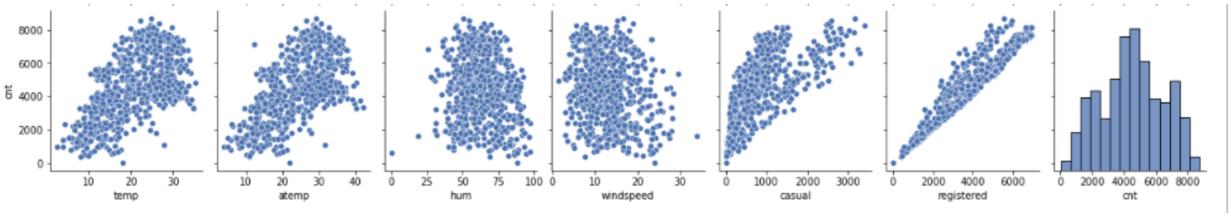
Where cnt is the target variable and the predictors are:
'temperature','windspeed','spring','winter','summer','Mist+Cloudy','year','holiday','Snow+Rain' ,’Tuesday'

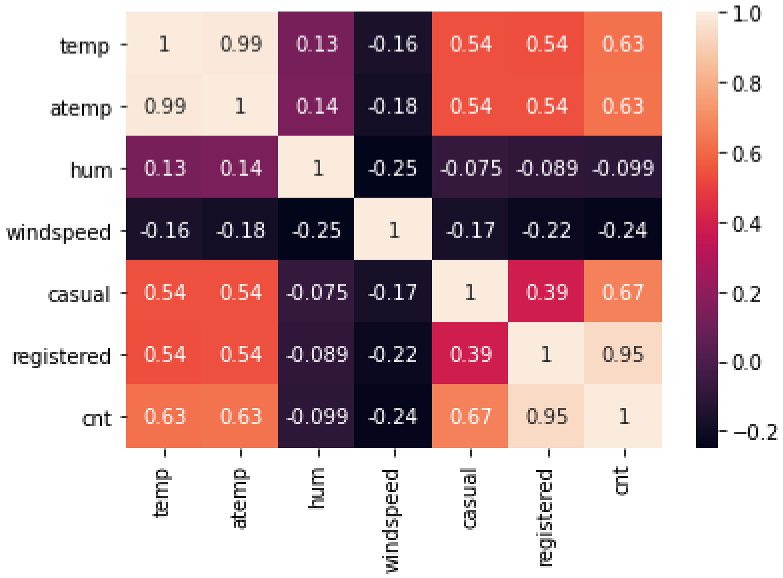
The model states that when one predictor for example temp changes by 1 unit, the target variable cnt changes by 4071.06 units when all the other predictors are kept constant.

Temperature, Winter, Summer, Year have positive coefficients which mean they are directly proportional to the target variable whereas Windspeed, Spring, Misty/Cloudy/Snow/Rain weather, Holidays and Tuesday are inversely proportional to the target variable as these values increase the target value decreases.

2. Why is it important to use drop_first = True during dummy variable creation?
Ans. When creating dummy variables in pandas using add_dummies(), it creates columns for all the levels in a particular column for example Season has four levels: Summer, Winter, Monsoon, Spring. If we use add_dummies() all these four levels will be converted to new columns with values 0/1. For minimizing the number of features we can drop one of these columns, as 0 is all other columns would indicate the dropped column.
Say we dropped column Spring using drop_first = True, Spring will be indicated by following scenario: Summer = 0, Winter = 0, Monsoon = 0. This will reduce the number of features to work with while building a model. Hence it is important to use drop_first = True during dummy variable creation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans. We can see that registered and casual variable which captures the number of registered and casual customer have a high correlation(0.95 and 0.67 respectively) but because casual+registered = cnt(target variable) we will drop these columns the next best correlation is with temp and atemp with a score of 0.63





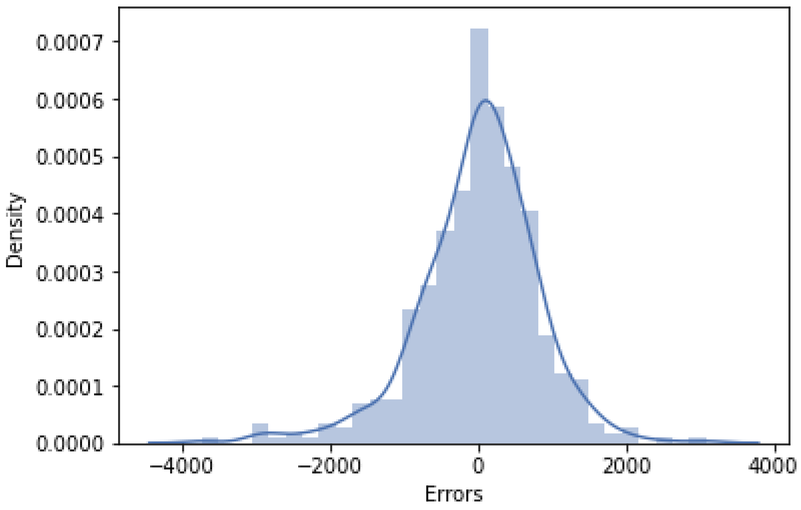
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

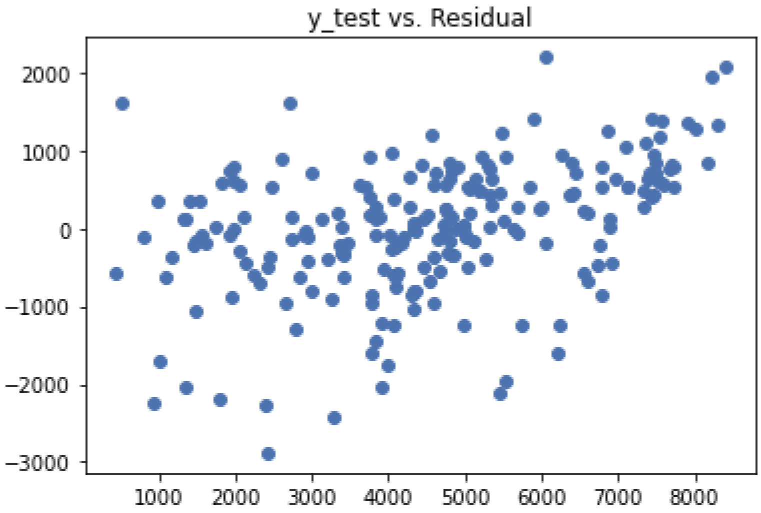
Ans. Following are the assumptions of Linear Regression:

- There is a linear relationship between target variable and predictors
- Error terms are normally distributed with mean at 0.
- Error terms are independent of each other as there is no visible pattern.
- Error terms have constant variance.

From the pair plot, it is evident that for few features there is a linear relationship with the target variable. We have done residual analysis and it is also evident that the error terms are normally distributed with mean at 0 and they have no visible pattern and have constant variance.

Distribution of Error Terms





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

Ans. The top three features are as follows as their absolute value of the coefficient is high:

1. Temperature
2. Snow+Rain
3. Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a statistical method that tries to show a relationship between variables. It looks at different data points and plots a trend line. A simple example of linear regression is finding that the cost of repairing a piece of machinery increases with time. There are a few assumptions for using linear regression.

- There is a linear relationship between target variable and predictors
- Error terms are normally distributed with mean at 0.
- Error terms are independent of each other as there is no visible pattern.
- Error terms have constant variance.

There are a few algorithms for linear regression

Ordinary least squares

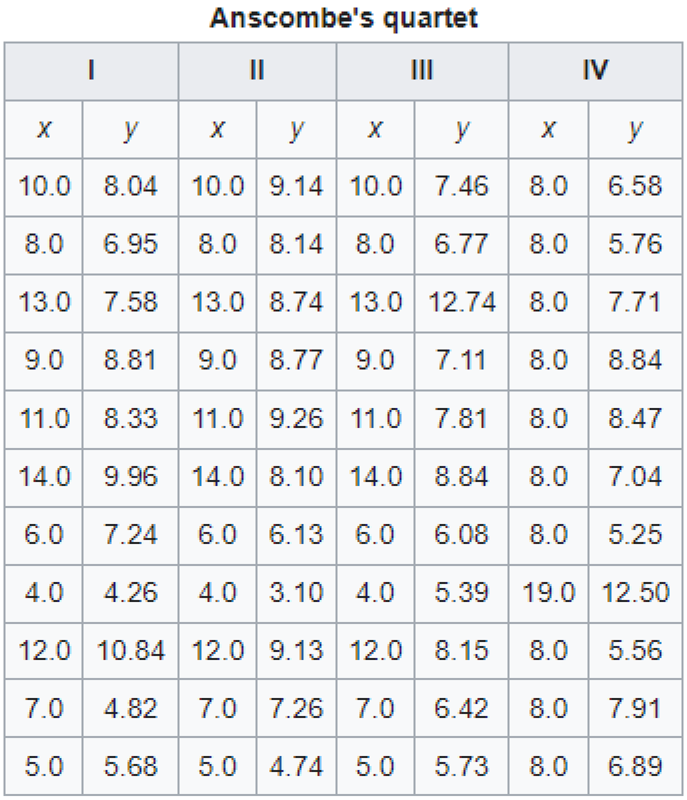
Ordinary least squares regression is a method to estimate the value of coefficients when there is more than one independent variable or input. It's one of the most common approaches for solving linear regression and is also known as a normal equation. This procedure tries to minimize the sum of the squared residuals. It treats data as a matrix and utilizes linear algebra operations to determine the optimal values for each coefficient. Of course, this method can be applied only if we have access to all data, and there should also be enough memory to fit the data

Gradient descent

Gradient descent is one of the easiest and commonly used methods to solve linear regression problems. It's useful when there are one or more inputs and involves optimizing the value of coefficients by minimizing the model's error iteratively. Gradient descent starts with random values for every coefficient. For every pair of input and output values, the sum of the squared errors is calculated. It uses a scale factor as the learning rate, and each coefficient is updated in the direction to minimize error. The process is repeated until no further improvements are possible or a minimum sum of squares is achieved. Gradient descent is helpful when there's a large dataset involving large numbers of rows and columns that won't fit in the memory.

2. Explain Anscombe's quartet in detail.

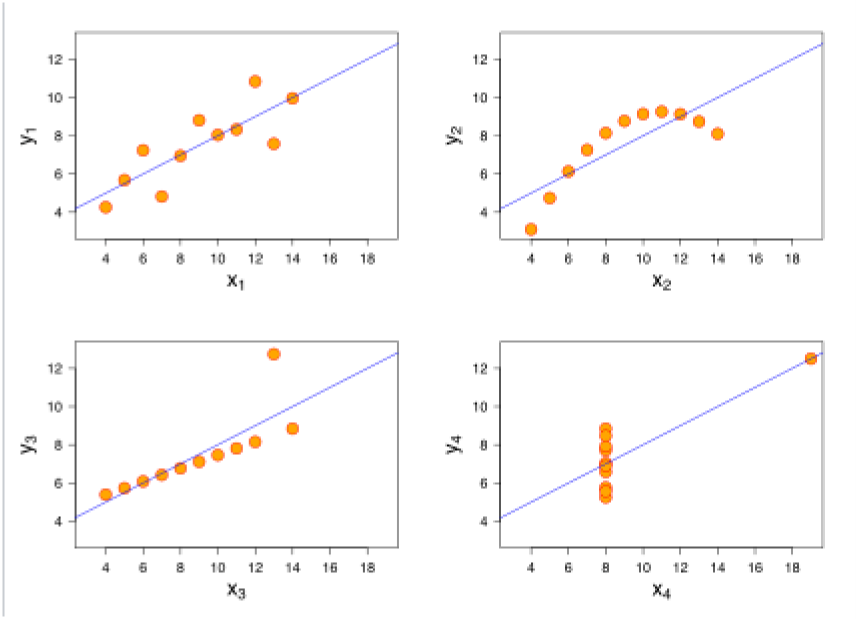
Ans. Anscombe's quartet suggests that we always plot the data rather than just checking the descriptive statistics such as mean, standard deviation, etc because even if the descriptive statistics are the same the data points when visualized can show a great difference. This happens generally because outliers affect descriptive statistics greatly. It was first stated by statistician Francis Anscombe in 1973. He formulated this approach to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

We take four completely independent datasets with 11 points. Even though their descriptive statistics suggest that the data points are equivalent, when we plot the graphs for these datasets, they are completely unrelated.



3. What is Pearson’s R?

Ans. Pearson’s R score indicates the correlation between variables. Its values lie between -1 to 1. 1 indicates a positive linear correlation and -1 indicates a negative linear correlation. According to the definition, Pearson’s R is the covariance of the two variables divided by the product of their standard deviations.

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

Ans. Scaling is the process of increasing or decreasing the magnitude according to a fixed ratio, you change the size but not the shape of the data. It is not mandatory to use feature scaling but it definitely is a good practice. It helps to handle disparities in units. During long processes, it definitely helps reduce computational expenses as the model building will be easier. The most common method of scaling is standardization, in this method, we centre the data, then we divide by the standard deviation to enforce that the standard deviation of the variable is one. Normalization most often refers to the process of constraining a variable to be between 0 and 1. This is also called as min-max scaling. The difference between Normalization and Standardization is that Normalization takes care of the extreme points as well by constraining all the data points between 0 and 1 whereas Standardization just shifts the dataset so that its mean is 0 and the standard deviation is 1.

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

Ans. Variation Inflation Factor(VIF) is given by the following equation 1/(1-R-squared). So when the R-squared value approaches to 1 the value to VIF approaches to infinity. This is the case of perfect correlation and indicates multicollinearity. For eg In the bike-sharing case study temp and atemp had a correlation of 0.99, if it were 1 these two features would have been a perfect correlation i.e. One variable is completely explained by another variable.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans. Quantile plot (Q-Q plot) graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. So in linear regression, we have an assumption that the error terms are normally distributed with a Q-Q plot we can visualise the same to validate. Q-Q plot can detect outliers, shifts in scale, location, symmetry, etc.