Linear Regression Subjective Questions

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?**

Answer: From the analysis of the categorical variables in the dataset we can infer that:

1. Demand for the year 2019 is larger than 2018, This can be because the popularity of this service is growing in people and since people are becoming more environment conscious, they find it a better option to opt for bike sharing
2. Demand for the bike is growing till the month of June. September has the highest demand, later on it decreases, Since this is the Fall season and the environment conditions are suitable during Fall season
3. Holidays have highest demand for bikes, as during holidays people may tend to find any physical activity as many have become health conscious.
4. The weathers it variable having Clear as the category has high demand, as the clear weather condition are favorable for bike rides

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

Answer: A variable having n labels are usually represented by n-1 dummy variables, because by default the value of one variable is assumed if others are not present in the data, we use drop_first = True, as it helps to reduce one extra column that are created during the dummy variable and hence the correlation created is also reduced

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

Answer:

1. Among all the variables present in the data, 'temp' and 'atemp' has the highest correlation with the Target variable, Since both the 'temp' and 'atemp' variables are correlated with eachother , we decided to drop any one of the variable to avoid the effect of multicollinearity which states that if two or more variables are highly correlated the impact of these variable remains same during model building or increases leading to the confusing as the model cannot determine which variable is responsible for the change in the r square value causing confusion.

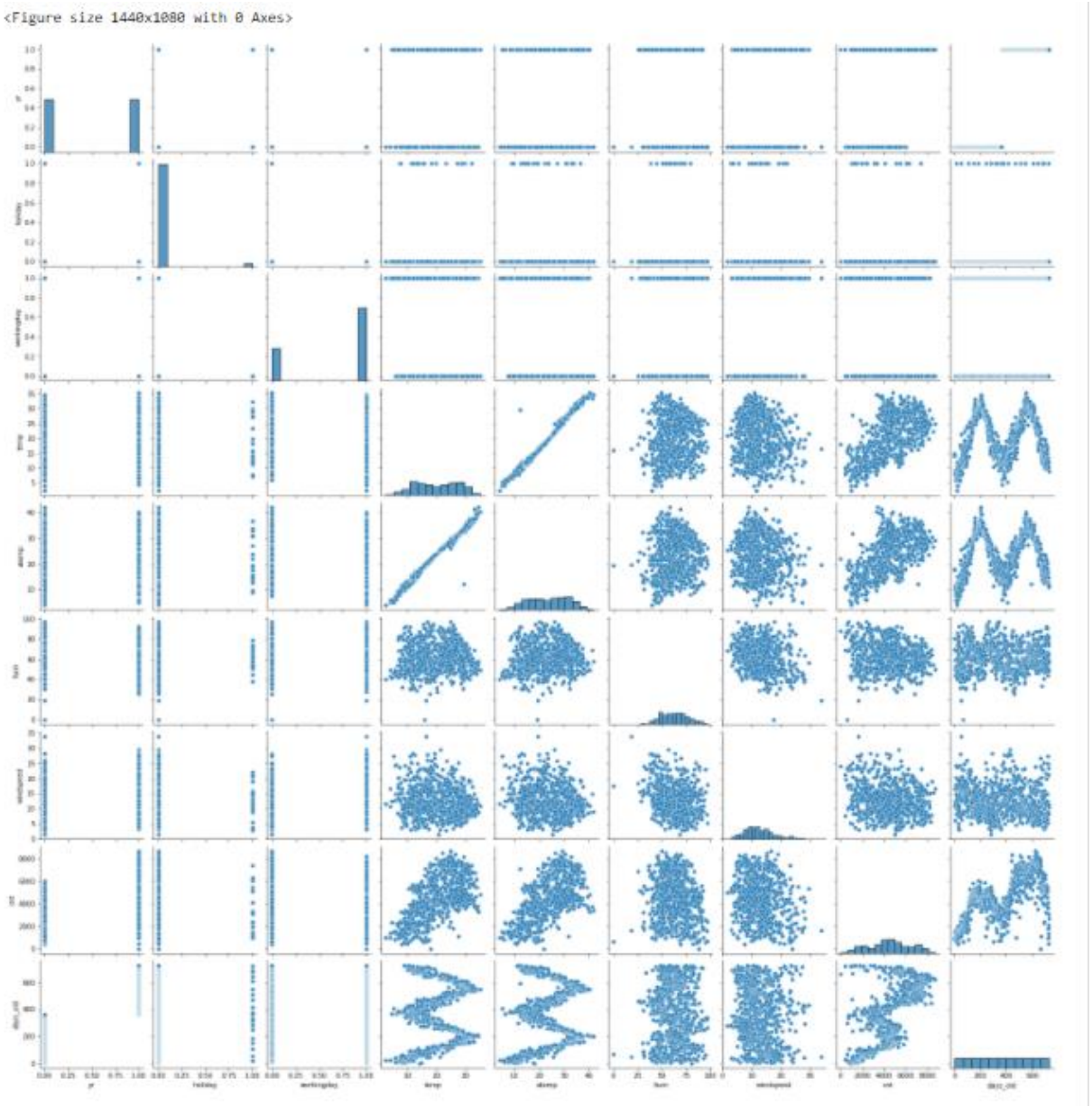
2. Thus, we decide to drop any one of the variables.

3. In this example we dropped the temp variable, thus 'atemp' is the numerical variable having highest correlation with the target variable 'cnt'.

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

Answer: Assumption for the Linear Regression that determine the model building are

1. The Dependent and Independent variable must have linear relationship:
   A pairplot shows the relationship between the target variable and all the other variables will give the verification for the linearity between the target and the other variables in the data.

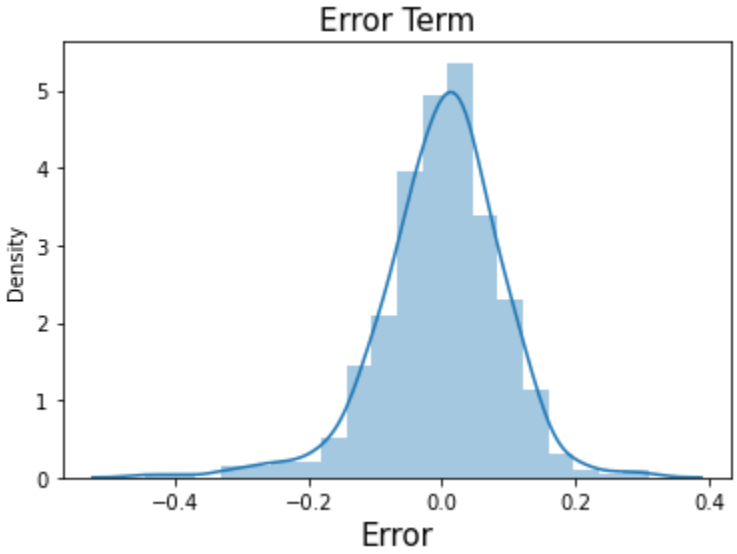


The lower right corner of the image shows the relation between the target and with other variable columns.

2. No Multicollinearity
   This can be seen by using heatmaps , when the variables are less. We can see the correlations and figure out the correlated variables. For the larger number of variables we can use the VIF(Variance Inflation Factor). If vif is 1 , we get very less multicollinearity, if VIF <5, we have moderate multicollinearity, if VIF > 5 , we have extreme multicollinearity.

   We can check the VIF for the variable and identify the highly correlated variables. We can decide to drop these feature or keep them depending upon the need of the requirement in business model

3. Residual must be distributed normally
   To validate these we used the distribution plot on the error terms and have seen the distribution of the data



The normally distributed image indicates the model is built correctly over the variables

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 three variables that significantly explains the demand for the shared bikes are atemp, year and weathersit_Light_Snow_rain_thunderstorm

1. atemp : This variable has the highest correlation with the target variable. It determines the feeling temperature in celsius. The temperature has the highest effect on the demand and inflation in the bike sharing service. During the high temperatures and sunny weather, it is less likely that the riders will opt for bike services as it gets exhausting riding bikes during sunny weather, similarly it's unlikely to ride the bike in extremely cold conditions. Thus, the favorable conditions show highest demand in bike riding
2. Year: It has been seen from the data that the demand for the bike riders rose in 2019 as opposed to 2018, this can be analyzed that the bike riding services might have got more popularity during year 2019, also people have become more aware about the environment conditions and thus has give preference to the fuel consumption free services which are economical as well as environment friendly
3. The third is the weathersit_Light_Snow_rain_thunderstorm , the weather situation where it has heavy rains or Ice pallets, thunderstorms, mist , snow , fog shows the negative effect on the bike riding service as its very obvious that people tend to stay at home of prefer bigger vehicle for traveling instead of bikes during such extreme weather conditions

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer: Linear Regression is the machine learning algorithm used extensively for predictive analysis. Machine learning is specifically the field of statistics which deal with the study of predictive modeling. It makes the model that provides the accurate predictions possible.

Linear Regression is the linear model that assumes the linear relationship between the input and the target variables.
A single variable input is termed as simple linear regression , whereas the multiple input variables are termed as the multiple linear regression.
The linear regression basicalay does:

a. Used to predict a variable and determine how accurate the prediction is done. Mostly used in time series or the future prediction analysis
b. Identifying the dominant variable amount the dataset, determining is magnitude and the beta estimation , ie whether it impacts positively or negatively

A simple linear regression form a model of:

$$y = B0 + B1*x$$

Where:

B0 - is the intercept of the regression line
B1 - is the slope or the coefficient of the variable

In higher dimension we have multiple variables , so the regression line is mentione as hyperplane.
If the coefficient becomes zero , it removes the influence of the input variable on the model and thus the prediction made becomes zero.

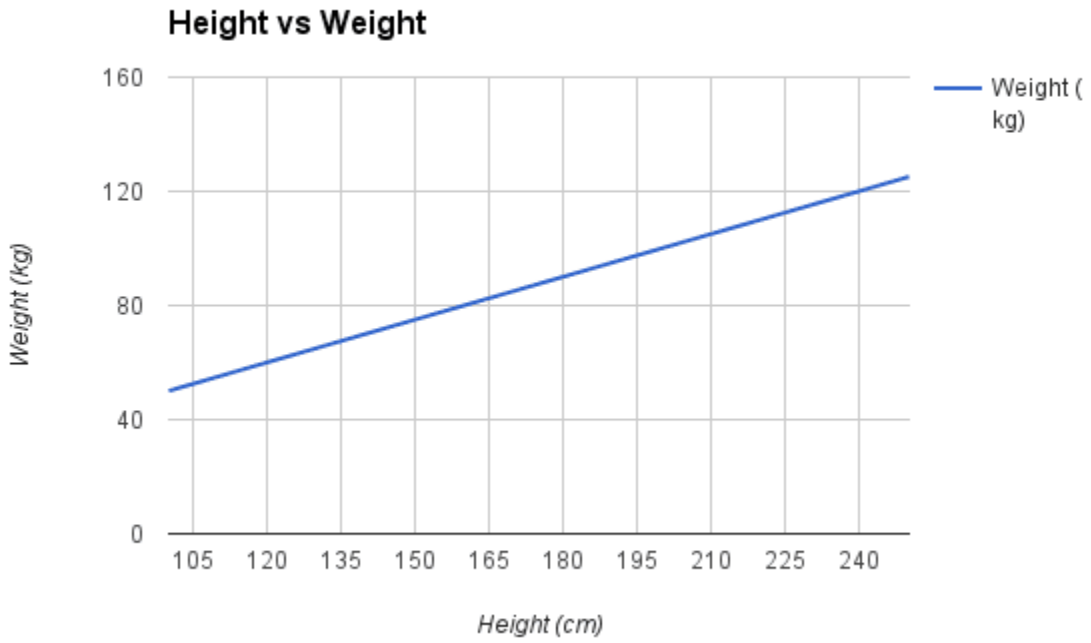
The major application of regression is in:

a. Determining the strength of the predictor
b. Forecasting

There are various type of linear regression

1. **Simple linear regression:**
    a. Gives the relation between a single input to the target variable
2. **Multiple linear regression:**
    a. Gives the relation between multiple variables with the output variable.
    b. Ordinary least square is the same method that gives the data as matrix and uses linear algebra to determine the optimal value of the coefficient. This method works better when it has least data to deal with
3. **Gradient Descent:**
    a. It starts with the random value of each coefficient and use the repeated process of sum of squares technique to improve the coefficients.
    b. In practice used=ful for the large dataset

4. **Regularization:**
    a. It tries to minimize the error in the model during training as well as reduce teh complexity of the model
    b. There are two types, Lasso Regression and the Ridge Regression



2. **Explain the Anscombe's quartet in detail.**

Answer:

1. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
2. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
3. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

4. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. o, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

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

Answer:

1. The Pearson correlation coefficient, also called Pearson's R, is a statistical calculation of the strength of two variables' relationships. In other words, it's a measurement of how dependent two variables are on one another. The Pearson product-moment correlation coefficient depicts the extent that a change in one variable affects another variable. This relationship is measured by calculating the slope of the variables' linear regression.
2. The value of Person r can only take values ranging from +1 to -1 (both values inclusive). If the value of r is zero, there is no correlation between the variables.
3. If the value of r is greater than zero, there is a positive or direct correlation between the variables. Thus, a decrease in first variable will result in a decrease in the second variable.
4. If the value of r is less than zero, there is a negative or inverse correlation. Thus, a decrease in the first variable will result in an increase in the second variable.
5. When plotted on a diagram, a positive correlation will see a line which slopes downwards from left to right and a negative correlation will see a line which slopes downwards from right to left.
6. Let's look at an example.

   Example:
   A classic case of two variables affecting one another is demand and supply in an economy when the price of the product and the quantity demanded and supplied is known. The values are represented using a simple linear regression. Pearson R shows that demand and supply have a positive correlation. As more consumers demand products, the amount suppliers are will to produce increases as well. The opposite is true with regard to price

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

Answer: Scaling: -

1. Scaling means that you transform your data to fit into a specific scale, like 0-100 or 0-1. You want to scale the data when you use methods based on measurements of the distance between data points, such as supporting vector machines and the k nearest neighbors. With these algorithms, a change of “1” in any numeric characteristic has the same importance.
2. For example, you could look at the prices of certain products in both yen and US dollars. One US dollar is worth about 100 yen, but if you don’t change your prices, methods like SVM or KNN will consider a 1 yen price difference as big as a difference of 1 US dollar! This does not correspond to our intuitions of the world. With currency, you can convert between currencies. But what if you look at something like height and weight? It’s not entirely clear how many pounds should equal an inch.
3. By scaling your variables, you can help compare different variables on an equal footing. To help solidify what scaling looks like, let’s look at an invented example:
4. scaling and normalization is that the terms are sometimes used interchangeably and, to make matters even more confusing, they are very similar! In both cases, you transform the values of numeric variables so that the transformed data points have specific useful properties. The difference is that:

    - when scaling, you change the range of your data, while
    - In normalization, you change the shape of the distribution of your data

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).

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

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

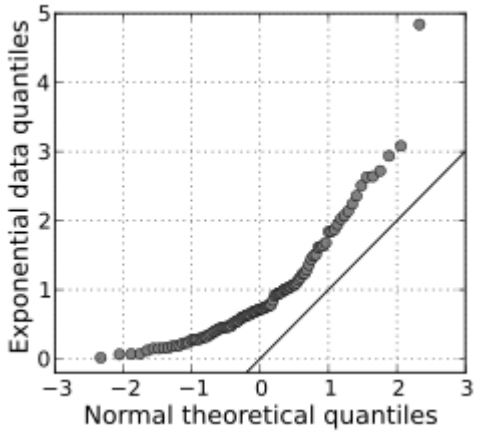
Answer:

1. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
2. Colinearity is the state where two variables are highly correlated and contain similiar information about the variance within a given dataset. To detect colinearity among variables, simply create a correlation matrix and find variables with large absolute values. In R use the corr function and in python this can by accomplished by using numpy's corrcoef function.
3. Multicolinearity on the other hand is more troublesome to detect because it emerges when three or more variables, which are highly correlated, are included within a model. To make matters worst multicolinearity can emerge even when isolated pairs of variables are not colinear.
4. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
5. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:

1. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
2. A Q Q plot showing the 45 degree reference line:



3. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
4. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.