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 – According to my studies in the current assignment I infer that the below pointers/variables should have the most effect on the dependent variable 'cnt'.

- Fall season attracts more hired rentals for bikes than in all other seasons.
- There is a steep rise in the count of bike rentals from the year 2018 to 2019.
- Most number of rentals are in the months from May to Sep. Highest peaks were observed during the plotting of the histogram.
- When its not an holiday people tend to hire lesser bikes so we can infer that the people hire slightly more bikes on holidays in lieu of spending personal time and enjoyment.
- Clear weather attracts more number of rentals rather than misty or light rainy scenarios.
- Low temp and high windspeed attracts less rentals as people don't want to ride bikes on hard/cold weather days.
- The working day/ non-working day seems to have no impact on the rental numbers.
- Thursday , Friday and the weekends attracted more number of rentals than other days of the week.

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

Answer – When creating dummy variables from categorical variables, the drop_first parameter is used to prevent multicollinearity, which can cause issues in regression analysis.

The following reasons support the above statement:-

1: Multicollinearity - Including all dummy variables can create multicollinearity, where two or more predictor variables are highly correlated. In the context of regression analysis, multicollinearity can cause issues like unstable coefficient estimates and inflated standard errors.

2: By setting the drop_first = True we drop the first level of each categorical variable when creating dummy variables. This effectively avoids multicollinearity because the dropped category serves as the reference category, and the information about it is captured implicitly in the intercept term of the regression model.

3: When converting categorical variables into dummy variables, you create m-1 dummy variables for a categorical variable with m amount of categories. For example, if we have a categorical variable "season" with three categories: Summer , Fall and Winter we would create two dummy variables: "Fall" and "Winter". The value of "Summer" is implicitly encoded when both dummy variables are 0.

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

Answer – According to the assignment the "Temp" variable has the highest correlation with the target variable.

---

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

Answer – The general validation of the assumptions of the linear regression is based on three factors as marked below:-

1: Normalcy of error terms – As we saw in the assignment that the error terms were evenly and normally distributed.

2: Multicollinearity check – As we saw and checked via plotting a heatmap for checking the collinearity amongst the variables. There exists almost 0 to very minimal collinearity.

3: Linearity – As we saw by plotting a CCPR plot that the dominant variable temp is showing a linearity pattern and the dataset values try to follow a linear pattern almost.

---

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

Answer – According the observations in the assignment the below three factors contributed the most for the demand of the shared bikes.

- "temp" – Being the most dominant variable
- "Sep"
- "winter"

General Subjective Questions

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

Answer – Linear regression is a statistical method used to model the relationship between a dependent variable (usually denoted as y) and one or more independent variables (usually denoted as x) by fitting a linear equation to observed data. It's one of the most commonly used approaches in predictive modeling and statistical analysis. Second commonly used approach is that of Logistic Regression.

Basic idea behind linear regression is to fit a best fitting line between the above mentioned two variables. Normally its denoted by the equation y = mx+c or y = *B0 + B1X which denotes a*

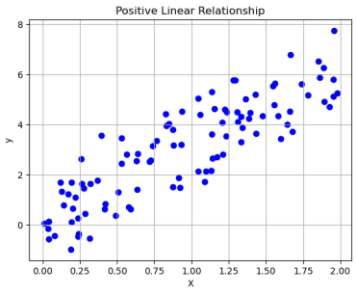
straight line.

Above equation *B0* is the intercept , *B1* is the slope , y is the dependent variable and x is the independent variable. Sometimes error term denoted by *e* is also added to predict the difference and error terms between y actual and y predicted.

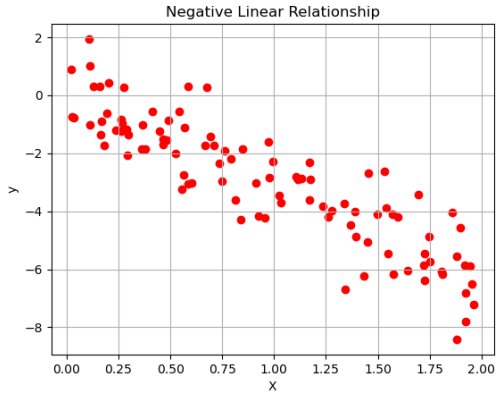
Example **–**

There can be 2 types of linear relationships. Positive and Negative.

1: Positive relationship – Here below we see that the data points follow an almost straight and positive relationship model. The positivity is highlighted by increase in Y values with increasing X values.



2: Negative relationship – Here below we can see the data points following an inverted pattern with negative slope. The negativity of the slope shows the decrease in the Y values with an increase in the X values.

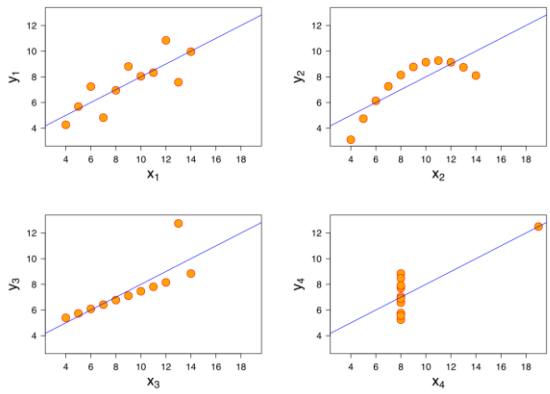


---

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

Answer – Anscombe's quartet is a famous example in statistics demonstrating the importance of visualizing data. It consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but have vastly different distributions and relationships when plotted. It comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

It is normally graphed or plotted before the actual statistical analysis begins and these 4 graphs summarizes the kind of touch points that one encounters in maximum of statistical problems.



---

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

Answer – Pearson's r, or Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1 and potentially has 3 values explained below:-

1 shows perfect linearity and perfect linear relationship.
0 shows no linear relationship
-1 shows a perfect negative linear relationship

It is normally used in statistics and data analysis for assessing the degree of association between two variables.

e.g. let's compare two values where in we take the speed of a vehicle and the average / litre of a vehicle while being driven on a straight road for a distance of 100 kms.

Speed of vehicle	Average/ litre of fuel
60	24km/litre
80	22 km/litre
100	19km/litre
120	16km/litre
140	10km/litre

To calculate Pearson's r, we first calculate the mean of X and Y, then calculate the covariance between X and Y, and finally, divide the covariance by the product of the standard deviations of X and Y. The formula for Pearson's r is:

X being the speed of vehicle , Xi being individual data point
Y being the average/ litre , Yi being the individual data point

To calculate r = sigma (Xi - Xmean)(Yi - Ymean) / sqrt(sigma (sq((Xi - Xmean))((Yi - Ymean))))

Final value of r is (**-0.9782**)
Hence, by the calculation of the r value or the Pearson value we can conclude that this is inversely proportional.

---

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

Answer – Scaling is a preprocessing technique used in machine learning and data analysis to standardize or normalize the features of a dataset. The main purpose of scaling is to bring all features to a similar scale or range. Scaling is important because many machine learning algorithms perform better when features are on a similar scale.

A: Normalization - Normalization scales the data to a fixed range, typically between 0 and 1. It is done by subtracting the minimum value of the feature and then dividing by the range of the feature (i.e., maximum value minus minimum value). Normalization is useful when the algorithm used expects normalized input values.

B: Standardization - Standardization scales the data so that it has a mean of 0 and a standard deviation of 1. It is done by subtracting the mean of the feature and then dividing by the standard deviation of the feature.

Mathematically, the standardized value of a feature x is calculated as:

X (standardized) = x – m / sd

Where m is the mean of the feature and sd is the standard deviation of the feature.

---

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

Answer – The Variance Inflation Factor (VIF) is a measure used to detect multi - collinearity in regression analysis. Multi - collinearity occurs when two or more independent variables in a regression model are highly correlated with each other.

Formula for calculating VIF (X) is – 1 / 1-R2 which implies 1 divided by 1-R square. Where x is one of the variables in question. Here R square is the coefficient that is determined by performing regression via other variables. Now its evident that if the value of 1-R2 is less then the value of VIF will increase. So if the R2 or coefficient of linearity between variables is less then the variable inflation factor is more.

Hence when the value of 1-R2 is 1 then it means there is perfect collinearity. Hence the value of Variable inflation factor is infinity. Normally this scenario is not a healthy scenario for plotting some inferences from the variable or while plotting collinearity.

---

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, short for quantile-quantile plot, is a graphical tool used in statistics and linear regression analysis to assess whether a given dataset follows a particular probability distribution, such as the normal distribution. In linear regression the Q-Q plot is used to check the normalcy of a

Working of a Q-Q plot :-

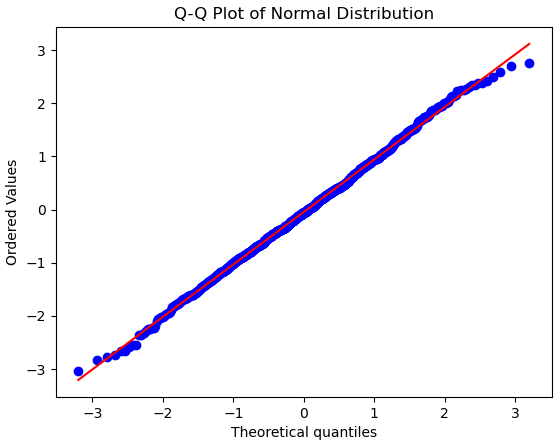
Normally this kind of plot is achieved via Scipy library.

Example - Below code shows the plotting of a Quarantile - Quarantile methology.

```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

# Generating random data from a normal distribution
np.random.seed(0)
data = np.random.normal(loc=0, scale=1, size=1000)

# Creating a Q-Q plot
stats.probplot(data, dist="norm", plot=plt)
plt.title('Q-Q Plot of Normal Distribution')
plt.xlabel('Theoretical quantiles')
plt.ylabel('Ordered Values')
plt.show()
```



- So I generated 1000 random data points from a normal distribution with mean 0 and standard deviation 1.

- Then I used stats.subplot from scipy.stats to create the Q-Q plot. Then specify the distribution we want to compare to (“dist = norm”) which is the normal distribution in this case.
- The plot the Q-Q plot using matplotlib.

The resulting plot will show the quantiles of our data against the quantiles of the theoretical normal distribution. If the points fall approximately along the straight line, it indicates that our data follows a normal distribution.