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)*

    - In spring month demand is less compared to other months.
    - Individual months Jun till Sep is the period where bike demand is high. January had the lowest demand.
    - 2019 there is demand increase compared to 2018.
    - Holidays vs non-holidays. Holidays had high demand.
    - The demand of bike is almost similar throughout the weekdays.
    - The bike demand is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall.
    - No data for Heavy Rain, Ice Pallets, Thunderstorm etc, SO no inference available. May be this can be future plan for the company.

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

Because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

looking at the pair plot temp & atemp variable has the highest of 0.63 correlation with target variable 'cnt'.

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

    - To verify the normality of error, distribution of residuals (vs) levels of the dependent variable using a QQ-plot and measure the divergence of the residuals from a normal distribution.
    - The independent variables was not be correlated. Multicollinearity was not considered.
    - The error terms were normally distributed.
    - No correlation between the residual (error) terms. Autocorrelation was managed.

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

Top 3 are: - year 2019; temp values (positive correlation) & light snow weather (negative correlation)

General Subjective Questions

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

It is basic form of machine learning where we train a model to predict the behaviour of data based on some variables. Linear regression (as the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated).

Mathematically, we can write a linear regression equation as:

y = a + bx

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

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

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 plot

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. Those 4 sets of 11 data-points are given below.

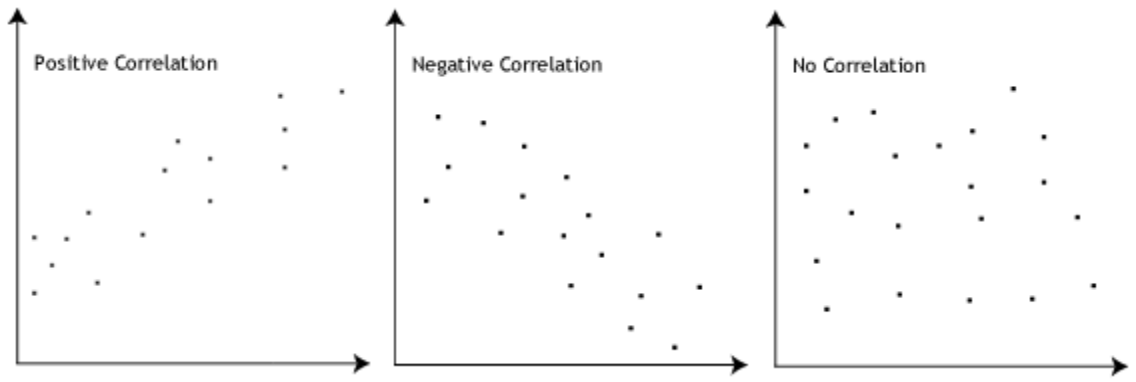
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

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

Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations. Thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

r =correlation coefficient
x_i =values of the x-variable in a sample
$\bar{x}$ =mean of the values of the x-variable
y_i =values of the y-variable in a sample
$\bar{y}$ =mean of the values of the y-variable

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0and1.*

  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**).*

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

- *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?*

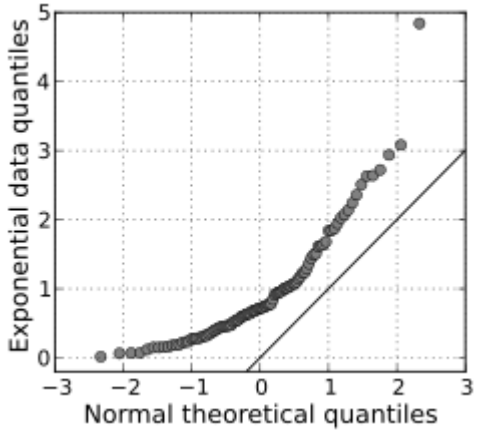
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

(3 marks)

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

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 instance, 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.

Please see snapshot below, a Q-Q plot showing the 45 degree reference line:



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.

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.

(3 marks)