Assignment-based Subjective Questions

Question 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: following points we could infer from data set:

- The demand bike increased in the year 2019 when compared with year 2018.

- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.

- Bike demand is less in holidays in comparison to when it is not a holiday.

- The demand of bike is almost similar throughout the weekdays.

- There is no significant change in bike demand with working day and non-working day.

- The bike demand is high when weather is clear and few clouds however demand is less in case of Light snow and light rainfall. We do not have any date for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog, so we cannot derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Kind of making dataset faster in case of big dataset. Ideally one should have to remove all null values from the columns from which we want derive dummy variables.

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

Answer: As per pair plot and heat map highest correlation with target variable 'cnt' is with 'atemp' i.e. actual temperature, which is 0.65. atemp is followed by temp with 0.64, then year (0.58).

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

Answer: For validation of linear regression model on train set we can check following:

- Value of **R^2** (**Percent of variance in observation**). It is calculated by (1-(RSS/TSS)). It varies from 0 to 1. If **R^2** value is close to +1. It means nearly all data points are explained by linear regression model and vice versa. Generally **r^2** value greater than **0.7** is considered to be a strong model.
- Value of **adj.R^2** (**Adjusted R square)**. It is square of correlation between score on dependent variable and value predicted by target variable after calculating number of predictor. It is calculated by {(1- R^2)(N-1)/N-p-1}, here R^2 is simple R square, N is total sample size and p is number of predictor. It shall always be positive and less than **R^2.** Adding more predictor or target variable tends to increase **R^2**. This is where **adj. R^2** comes in picture to explain data points by model.
- All p values should be below 0.5.
- Check for residual analysis. Distributions of residuals should be Gaussian and preferably non-skewed.

Question 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: Based on final model top three features contributing significantly towards explaining the demand are:

- Actual Temperature (0.5695). More the temperature more shall be target variable.
- weathersit_3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.2553). If weather is under third condition, It would negatively affect target variable.
- Year (0.2283). Over a period of time bike sharing idea would become trend. Hence with passing of years target variable shall increase more.

General Subjective Questions

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

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variable. It is mostly used for finding relationship between variable and forecasting. The whole idea is to get a best fit in line against all data points usually with RSS or OLS method. Linear regression best fit line is represented by equation:

y = c+ m1X1+m2X2+ ….. + mnXn+e

here y is value of target variable

c is intercept

e is error

m1 is coefficient if variable x1 and so on up to n

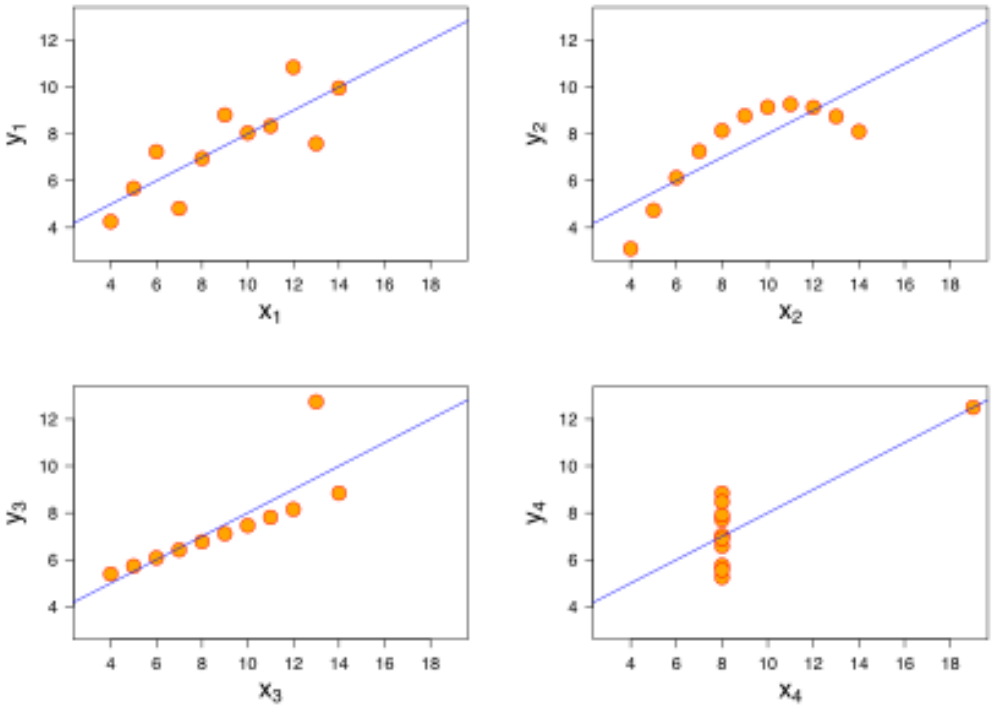
- Following are assumptions in linear regression.

- It is assumed that target and dependent variables are in linear relationship.
- Assumptions about residuals:
  - Residuals are in Gaussian distribution
  - Residuals have zero mean value
  - Residual terms have same variance. i.e. homogeneity in error terms.
  - Pair wise covariance is zero. i.e residual terms are independent to each other.
- Assumptions about Estimator:
  - Independent variables are measured without error.
  - Independent variables linearly independent to each other. There is no multicollinearity in data.

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

Answer: Anscombe's quartet can be defined as group of four data sets which are nearly identical in simple descriptive statistical details, but there are peculiarities in data set that fool he regression model if build. They have very different distributions and appear differently when plotted in scatter plots. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it. The effect of outliers and other influential observations on statistical properties. He proved power of data visualization.

Anscombe's quartet							
A		B		C		D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.1	10	7.46	8	6.6
8	6.95	8	8.1	8	6.77	8	5.8
13	7.58	13	8.7	13	12.7	8	7.7
9	8.81	9	8.8	9	7.11	8	8.8
11	8.33	11	9.3	11	7.81	8	8.5
14	9.96	14	8.1	14	8.84	8	7
6	7.24	6	6.1	6	6.08	8	5.3
4	4.26	4	3.1	4	5.39	19	13
12	10.8	12	9.1	12	8.15	8	5.6
7	4.82	7	7.3	7	6.42	8	7.9
5	5.68	5	4.7	5	5.73	8	6.9



FOUR GRAPHICAL PLOTS OF DATA SETS ARE DIFFERENT FROM EACH OTHER

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

Answer: In statistics, the Pearson's correlation coefficient (R) is the product-moment correlation coefficient, the bivariate correlation is a measure of linear correlation between two data set. It is calculated by following formula as mentioned below:

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

Where,

r = Pearson Correlation Coefficient

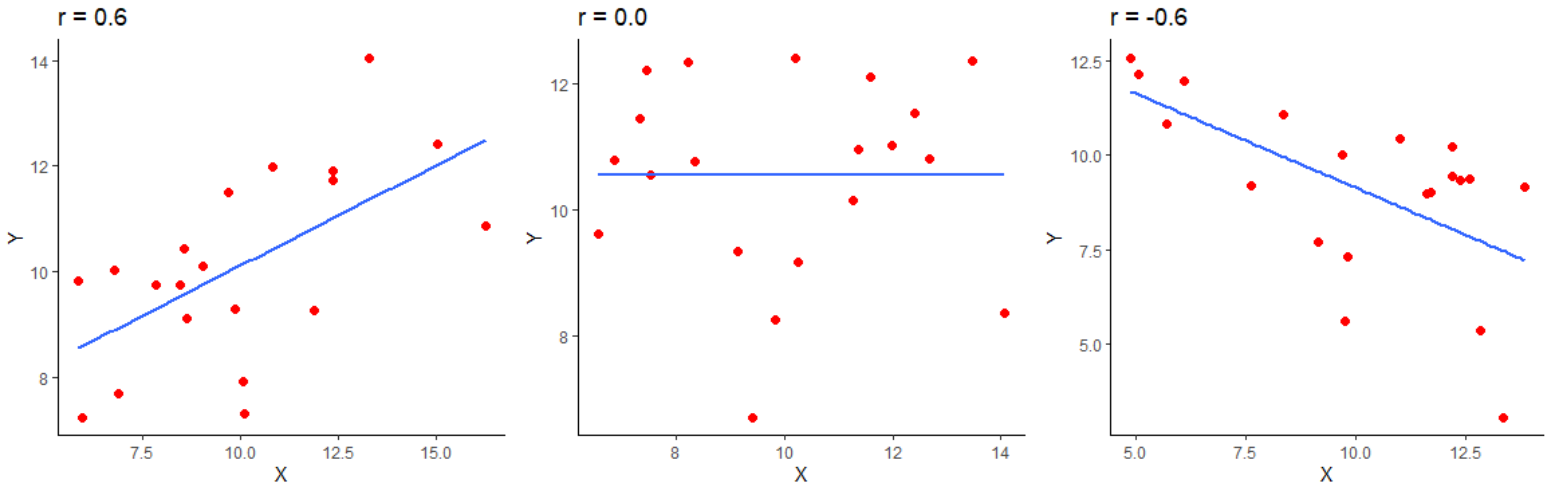
x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

- The value of R ranges from -1 to +1.
- The sign of R indicates direction.
  - Positive sign means direct relation between two datasets and vice versa.
- The value of R indicates strength of correlation.
  - Value close to one shows strong correlation and close to zero is weaker one.



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

Answer: Feature scaling is method to normalize the range of independent variable or features of dataset with in particular range. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. It also helps in speeding up calculation for 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 deal with this issue, scaling is done to bring all variables at same level of magnitude. It is important to note that scaling just affects the coefficients and not any other parameter as t-statistic, F-statistic, p- values and R-squared, etc.

Scaling is of two types:

- **Normalized Scaling:** It is a technique in which values are shifted and rescale so that they end up ranging between zero and one. It is also called as Min-Max scaling.
    - Formula for Normalized Scaling is X' =((X-Xmin)/(Xmax-Xmin))
    - When X is min value, numerator shall be zero. Hence X' shall be zero.
    - When X is max, then X' shall be 1.
    - If X is in between min and maximum, value of X' shall be in between 0 and 1 respectively.
    - It is used usually when data points do not follow Gaussian curve.
    - It supports greater overall database organization
    - Reduce redundant data
    - Data consistency with in the database
    - A much more flexible database design
- **Standardized Scaling**: It is another technique of feature scaling where values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and resultant distribution has a unit standard deviation.
    - It is calculated by X' = ((X-μ)/σ).
    - Here μ is mean of feature
    - And σ is standard deviation of feature values.
    - It is used when data points are distributed on Gaussian curve.
    - It is useful where data have negative value.

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

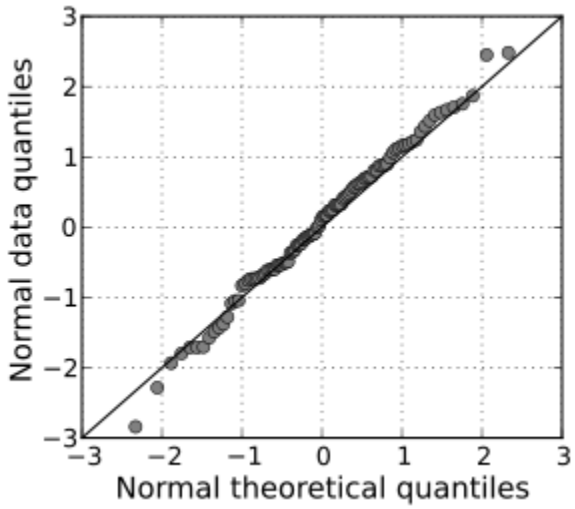
Answer: If there is a perfect correlation, then variance inflation factor shall be come infinite. In case of perfect correlation we shall have $\mathbf{R^2}$ as one. Which lead to **VIF** = (1/(1- $\mathbf{R^2}$)) as infinite. To solve this problem we need to drop one of the variable from dataset which is causing this perfect multicollinearity.

An infinite VIF indicates that the corresponding variable may be expressed exactly by a linear combination of other variable (which show an infinite VIF as well).

As a referral method to identify multicollinearity between variables we calculate VIF. IF VIF is 1 there is absolutely no correlation between two variables. Ideally VIF between one to four is considered good for variable to be used for model building. If VIF is between 5 to 10, variable should be dropped out while model building.

Question 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 (Quantile-Quantile plot) 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° 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.

If the two distributions being compared are similar, the Q-Q plot will approximately lie on the line y=x. if the distribution is 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 distribution.

A Q-Q plot is used to compare shapes of distributions, providing a graphical view of how properties such as location, scale and skewness are similar or different in two distribution.