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: There are 7 categorical variables in the bike sharing dataset

   - **Season:** ~60 % people preferred rides during fall and summer whereas the least preferred season for riding is spring (14%)
   - **Year:** 1.6X growth in 2019 compared to 2018
   - **Month:**
     - Highest bike engagement happened during mid of the year (May to Oct ~60 %)
     - In Jan and Feb, we can suggest to service the bikes, without much impacting the business Casual Rides
   - **Holiday:** 97 % of the total rental bikes are during non-holiday period
   - **Weekday:** Most of all the day's people prefer riding
   - **Working Day:** ~3/4th of the bikes are due to work day
   - **Weathersit:** 70 % of the riders prefer to ride in Clear weather (Few clouds and Partly cloudy)

   Addition to the above:

   - **Season v/s Year:**
     - In all the seasons of 2019 the number of bikes are more compared to last year
     - Lowest bikes recorded in the 2019-winter season
   - **Temperature Bins:** ~40 % of the rides are during medium feeling temperature (20-30 C)
   - **Humidity Bins:** 65 % of bike sharing during (50-75) humidity range
   - **Wind Speed Bins:** 90+ % of bike sharing in low wind speed (<20)

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

   ANSWER: Dummy variable creation helps us to represent the categorical variables into numeric.

   For Ex: In the season variable we have 4 categories (1: spring, 2: summer, 3: fall, 4: winter). Dummy variable creates a 4 variables with 1 & 0, if we won't specify the drop_first=True. If we have all zeros in spring, summer and fall, it represents it's a winter season. No need of extra variable to represent.

   If we use drop_first=True while creating dummies, it avoids the multi-collinearity between dummies.
   If there are n categories, we need to create n-1 dummy variables

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

   ANSWER: Highest correlation is among temp and atemp variables. It indicated the presence of multi-collinearity. Both these variables can't be used for model building, need to drop one variable depending on business requirement, VIF, p-values w.r.t. other variables

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

   ANSWER:

   - Very low multi-collinearity: Using VIF, considering variables within permissible range (<5)
   - Residuals or error must be normally distributed - Plotting
   - No visible patter between residuals and dependent or independent variable - plotting
   - All the independent variables in the model must be significant in nature; this can be identified by p-value based on null hypothesis (Co-efficient of independent variables are equal to zero)
   - F-statistics and probability of F-statistics indicates the overall performance of the model
     - F-Statistics: 320.7
     - Prob(F-stat): 7.75e-192

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

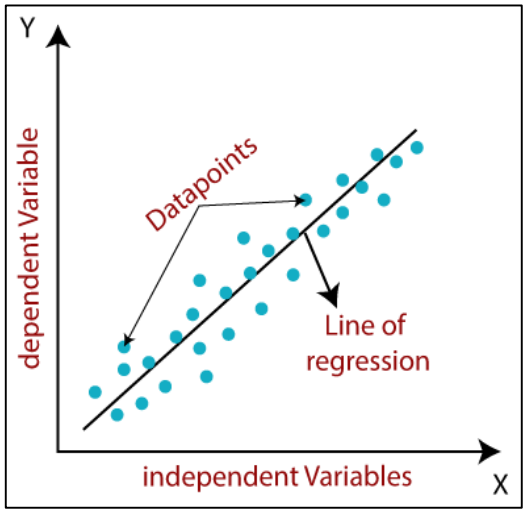
ANSWER: atemp, weathersit_3 and yr_2019 variables

- **atemp**: A unit increase in feeling temp variable, increases the bike rides by 0.378637 units.
- **weathersit_3**: A unit increase in weathersit_3 variable, increases the bike rides by 0.284695 units.
  - weathersit_3: (Light Snow, Light Rain, Thunderstorm, Scattered clouds)
- **yr_2019**: A unit increase in yr_2019 variable, increases the bike rides by 0.253157 units.

General Subjective Questions

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

ANSWER:



Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Two types:

1. Simple Linear Regression: Single independent variable
2. Multiple Linear regression: More than one

Mathematically,

$y = a_0 + a_1 x + \varepsilon$ where,

Y = Dependent Variable; X = Independent Variable

a0 = Intercept of the line; a1 =Linear regression

ε = random error

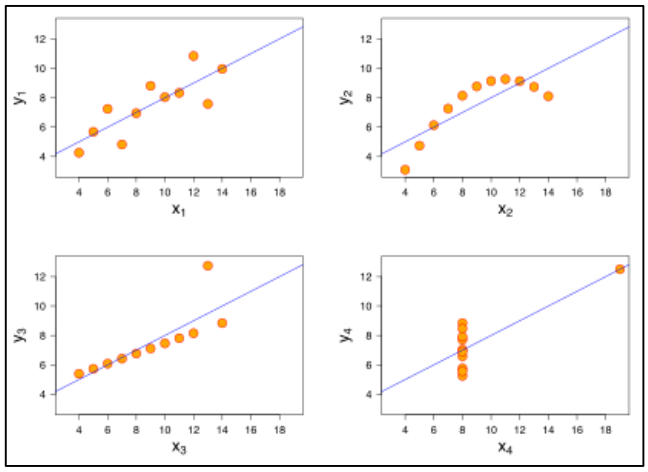
The cost function helps us to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. By using this MSE as a cost function we are going to change the values of a_0 and a_1 such that the MSE value settles at the minima.

Assumptions:

- Linear relationship between dependent and independent variables
- Error terms are normally distributed with mean zero
- Error terms are independent of each other – No visible pattern
- Error terms are constant variance – Homoscedasticity

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

ANSWER:



Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

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.

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 reparability of the data, etc. Also, the Linear Regression can be only being considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.

The four datasets can be described as:

- Dataset 1: Fits the linear regression model pretty well.
- Dataset 2: Could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: Outliers involved in the dataset which cannot be handled by linear regression model.
- Dataset 4: Outliers involved in the dataset which cannot be handled by linear regression model

Conclusion:

We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

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

ANSWER: In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1 and +1.

It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

$$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

Pearson correlation	**Greater than 0.5**	**Between 0.3 and 0.5**	**Between 0 and 0.3**	**0**	**Between 0 and – 0.3**	**Between 0.3 and – 0.5**	**Less than – 0.5**
Strength	Strong	Moderate	Weak	None	Weak	Moderate	Strong
Direction	Positive	Positive	Positive	None	Negative	Negative	Negative

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

ANSWER: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also 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.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Sl No.	Normalisation	Standardisation
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is an often called as Scaling Normalization	It is an often called as Z-Score Normalization.

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

ANSWER: 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 multi-collinearity.

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).

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 multi-collinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multi-collinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANSWER: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

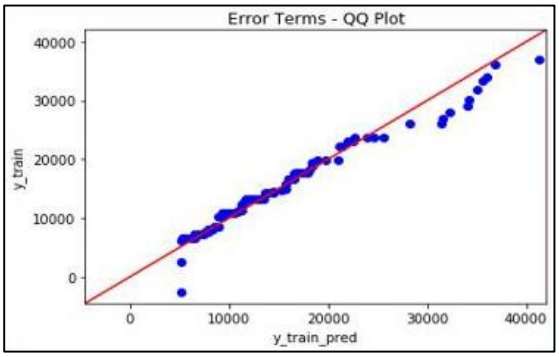
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

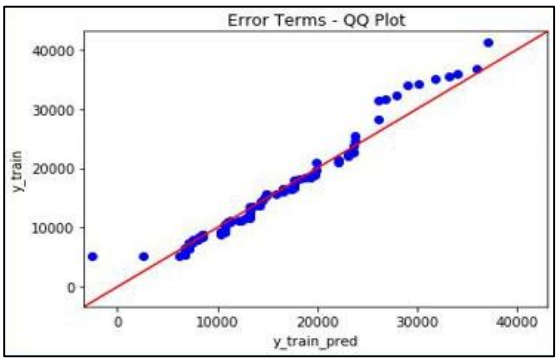
Advantages: If two data sets are-

- Come from populations with a common distribution
- Have common location and scale
- Have similar distributional shapes
- Have similar tail behaviour

Interpretation: *A* q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degrees from x –axis





- Y-values < X-values: If y-quantiles are lower than the x-quantiles.

- X-values < Y-values: If x-quantiles are lower than the y-quantiles.

- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degrees from x -axis