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?

Solution:

- "mnth": It can be seen that the months, January, February and March see the lowest sales. April onwards the sales start to pick up, with the peak in the months of June to October. Sales start dipping from November again.
- "holiday": As expected, there are more sales on a holiday.
- "weekday": The sales pretty much remain constant with some lows in the middle of the week.
- "workingday": Sales are pretty much the same throughout.
- "weathersit": If the weather is "Clear, Few clouds, Partly cloudy, Partly cloudy", the sales are more. A bad weather definitely does affect bike rentals
- "season": Summer and Fall see the highest number of rentals, while spring shows the lowest number of rentals.
- "yr": 2019 had more sales, given the COVID situation, probably.

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

Solution:

- When modeling a problem as a linear regression, we must make sure that the columns are linearly independent of each other so that the space spanned by W.T.dot(X), where *W* is the coefficient matrix and *X* is the input is equal to the dimension of the input.
- By using drop_first=True, we ensure that an extra redundant column is not created when we one hot encode a particular categorical column. This avoids multicollinearity, one of the main assumptions for a linear model to work.

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

Solution:

- “temp", “atemp” has more of linear relationship with “cnt” than any other variables

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

Solution:

- Plotted pdf of residuals and ensured that it has a normal distribution
- Plotting a pairplot showed a linear relationship between X and y

- VIF was calculated on each column on the final model deployed. Verified that all the columns had a VIF < 5

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

Solution:

- Weather, year, month and temperature are the key indicators for the number of bike rentals.
    - A bad weather impacts the sales for sure. Weathersit = 3 has a coefficient of –2581.97
    - The year 2019 had a lot of sales with a coefficient of +2035.43
    - From our bivariate analysis as well, we can see that month is a good indicator, with the peak sales in the months of June to October. From the model we see, September has a coefficient of +1667.7
    - This is followed by some more month related variables and then temperature with a coefficient of +819.1

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- The objective of linear regression is to be able to predict the value of a continuous (dependent) variable, y by a linear equation of input variable(s), X. Here we ensure that the input variables are linearly independent of each other and that the vector y actually lies in the span of the input space.
- Besides this, for linear regression to work, the error terms obtained must have a normal distribution.
- If there are N samples and d is the dimensionality of input vector, X. then W is a column vector of size, dX1, which are the tunable parameters of the model. X = NXd, then y_hat(predicted output) = X.dot(W) --> NX1
- We minimize the mean squared error (Least mean square) given by:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

makes result quadratic

y_i: true y; $\hat{y}_i$: estimate of y

$\frac{1}{N}$: average over all results

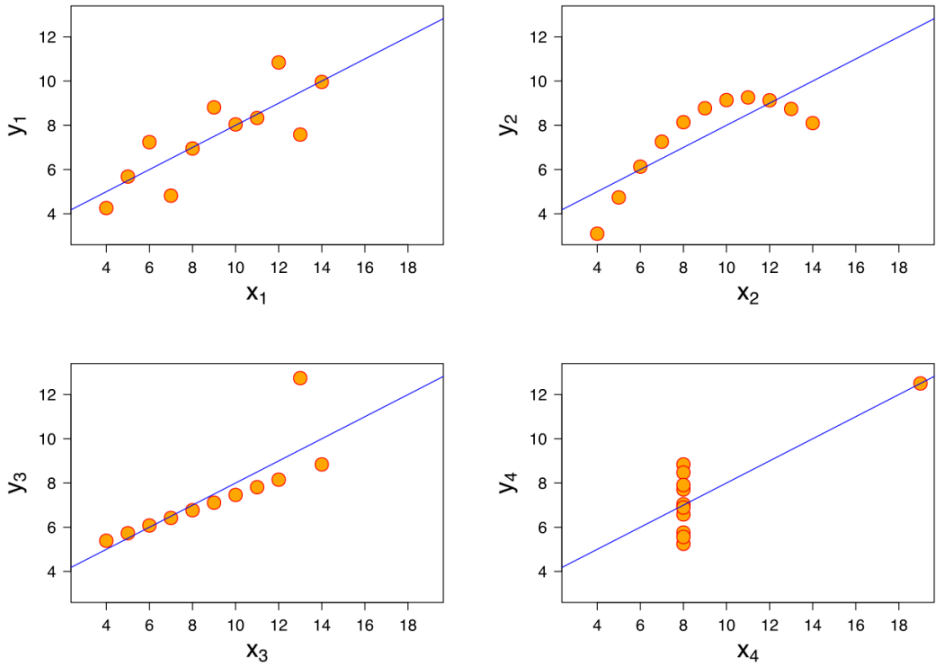
with respect to W, to find the optimal W.

- There are two ways to do this: One is a closed solution which involves a pseudo-inverse of X. The other way is to do online update. Here, we take the gradient cost function and go in the opposite direction by a small step size, mu.
- Based on this step size, we slowly approach the minima. The weight update equation is given by: W = W +mu*error*X, in vector form.
- The closed form solution is not feasible at times as it is computationally intensive and we cannot do an online update either. In many problems like in reinforcement learning or channel estimation, etc. we want to update the weights in an iterative manner and reach convergence.

2. Explain the Anscombe's quartet in detail.

Solution:

- Say you have two datasets, y1 = f(x1) and y2 = f(x2), having the same summary statistics like mean, variance, Correlation between x and y and the same linear regression line. We might think that when plotted their distributions might look the same.
- But that's not always the case, therefore it is important to always plot the data and not just rely on summary statistics when comparing two datasets.
- Additionally, Anscombe's Quartet also warns of the dangers of outliers in data sets.
- Take the example of the following four plots:



- It is clear that the second graph should not be modelled as a linear regression at all. But the top left plot should.
- Therefore, before analysis, it is important to graph the data.

3. What is Pearson's R?

Solution:

- It is the most common way of calculating correlation between two vectors. The formula for Pearson's R coefficient for two vectors X and Y is given as:

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- A –1 implies a perfect inverse correlation, whereas a +1 indicates that the two quantities are directly proportional to each other.
- In Pandas, the corr(), by default implements the Pearson's R correlation. Unless specified otherwise.
- During analysis, we can say that the correlation is weak, if the coefficient is between 0.3 and 0.5, while 0.5 to 0.7 implies a moderate correlation and greater than 0.7 implies a strong correlation.

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

Solution:

- Machine learning and deep learning models use gradient descent for updating the weights(at each layer for deep neural networks). The weight update equation as pointed out before in vector form is given by: W = W +mu*error*X
- Therefore, the update is directly proportional to the input, X. The step size will be drastically different if features are not on the same scale. To ensure the update is smooth towards the minima, we must bring them to the scale. This is where feature scaling comes in.
- Normalization also known as Min max scaling is given by X' = (X-Xmin)/(Xmax-Xmin), whereas standard scaling, also known as Z-scoring is given as: X' = (X – mu)/std_dev
- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This is the preferred scaling for when we using neural networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range

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

Solution:

- Variance Inflation Factor (VIF) is given as 1/(1-corr_coeff), where corr_coeff is the correlation coefficient between the input vectors. It is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
- If the input variables are not linearly independent, and in fact if it turns out that it is perfectly correlated then corr_coeff = 1. In this case, VIF will show a value of infinity.
-

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

Solution:

- Q-Q plots help us to identify whether two data sets are coming from the same population with same distributions. It is a graphical tool to help assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
- In linear regression, we can assess if training and test data sets are coming from the same distributions.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.
- If all points of quantiles lie on or close to straight line at an angle of 45 degree from x –axis, we can conclude that the two data sets are coming from a similar distribution.
- If all the points of quantiles lies away from the straight line at an angle of 45 degrees from x-axis, we can say that the two data sets are coming from a different distribution.