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)**

    - cnt of bike sharing is highest in Fall season followed by summer, winter and lastly spring
    - cnt of users increased in 1 i.e yr 2019 compared to in 0 i.e 2018
    - cnt of users is comparitively less on holidays
    - cnt of users sharing bike is highest when weather is clear i.e. 1 and lowest when weather is rainy / snowy i.e. 3

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one.
   pd.get_dummies() will by default create n columns for n levels. **drop_first=True** drops the first column thus creating n-1 new columns each indicating whether that level exists or not

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   From the pair-plot among the numerical variables we can say that "registered" (users) variable have highest correlation with the target variable i.e. "cnt"

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

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

General Subjective Questions

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

   Linear Regression algorithm is one of the to do algorithms in our day to day data science project dealing essentially with continuous form of data. In this algorithm we use the high school mathematics formula of y=mx+c and emphasize on it using either calculus or advanced matrix operations to find the weight vectors. Here the y vector is the label matrix and the x is that of input or feature matrix while m signifies the weight vector and c the intercept vector generally called a dummy vector.

In this algorithm there are two broads based at the beginning one of a analytical approach and another of a gradient descent approach.

In an analytical approach it's assumed that the dataset has a perfectly good linear relationship and can be solved using analytical approach where the weight vector is (Transpose(X).X)inverse.(transpose(X).Y) .

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

But this assumes that it fits the data perfectly while for a dataset which doesn't have a perfect fit the gradient descent approach is more apt where the derivative of the loss is considered for optimization of the weight vector i.e for successful update.In the gradient descent approach there are again two types one without regularization and one with i.e adding a bias vector with a regularization constant to balance overfitting in case.

There are three types of weight updates methods i.e full batch gradient descent approach in which the entire data is fit in one go and weight updates for all is carried out only once,another is called the mini batch where batches of a particular size parameters are made and iteratively one by one they are fit and updates and the last is stochastic where the size is that of 1 i.e one data is trained at a time and updated once and then moves to the next data row for optimization.

The loss formula for linear regression i.e the where hw signifies the model function and x is the feature vector, y is the label vector.

$$\begin{aligned} J(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}\left(h_{\mathbf{w}}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2 \\ &= \frac{1}{2}\sum_{i=1}^{n}\left((w_0 + w_1 x_1^{(i)}) - y^{(i)}\right)^2 \end{aligned}$$

The loss function is further differentiated taking partial derivatives as below.

$$
\begin{aligned}
\frac{\partial J(\mathbf{w})}{\partial w_0} &= \sum_{i=1}^{n}(w_0 + w_1 x_1^{(i)} + \ldots + w_m x_m^{(i)} - y^{(i)})x_0^{(i)} \\
&= \sum_{i=1}^{n}(w_0 + w_1 x_1^{(i)} + \ldots + w_m x_m^{(i)} - y^{(i)})1 \\
\frac{\partial J(\mathbf{w})}{\partial w_1} &= \sum_{i=1}^{n}(w_0 + w_1 x_1^{(i)} + \ldots + w_m x_m^{(i)} - y^{(i)})x_1^{(i)} \\
&\vdots \\
\frac{\partial J(\mathbf{w})}{\partial w_m} &= \sum_{i=1}^{n}(w_0 + w_1 x_1^{(i)} + \ldots + w_m x_m^{(i)} - y^{(i)})x_m^{(i)}
\end{aligned}
$$

The weight vectors are further optimized using the partial derivatives as below.

$$
\begin{aligned}
w_0(\text{new}) &:= w_0 - \alpha\frac{\partial J(\mathbf{w})}{\partial w_0} \\
w_1(\text{new}) &:= w_1 - \alpha\frac{\partial J(\mathbf{w})}{\partial w_1} \\
&\vdots \\
w_m(\text{new}) &:= w_m - \alpha\frac{\partial J(\mathbf{w})}{\partial w_m}
\end{aligned}
$$

Here alpha is the learning rate i.e the rate at which the partial derivative is given weightage to… if too high the global minima may be overshot i.e it basically controls how fast or how slow the global mimima can be reached.

The weight vectors are updated as below and the optimization procedures are those of Stochastic,Mini batch and full.

Finally the evaluations is carried out using the below formula.

$$\text{RMSE} = \sqrt{\frac{1}{n}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})}$$

Here the final loss of the model is calculated post optimization on the test data set and this measures how well our model has performed.

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

Anscombe's Quartet demonstrates the importance of data visualization before analysing it with statistical properties.

Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.

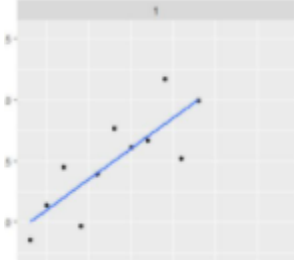
Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties such as mean, standard deviation and variance and all four differ with each other graphically. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

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

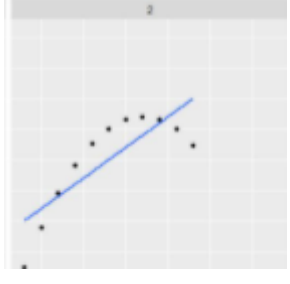
Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

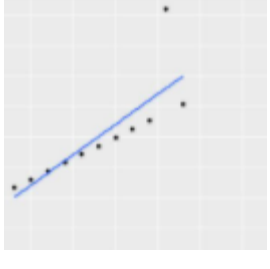
In the first one there is a linear relationship between X and Y.



In the second oneHere there is a non linear relationship.



In the third one There is a linear relationship here as well but with the exception of an outlier.



Finally, the fourth one shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

   A Pearson's correlation coefficient is a measure of the correlation between 2 linearly related variables. It ranges from -1 to +1. -1 indicates a highly negative correlation whereas +1 indicates a perfectly positive correlation. Generally, a measure of less than 0.5 could be taken to indicate a low association between the variables. It is computed as Covariance/product of the standard deviations of the variables.

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

   It is likely that values in a dataset belong to a large range and/or display significant variance. In order to standardize such data, they are reduced to a common scale using scaling. Normalised scaling reduces values to a range between 0 and 1, whereas standardized scaling ensures that the data is distributed around 'mean' 0 and has a variance of 1. Scaling is commonly used in machine learning to reduce the time taken for convergence/computations, and also to eliminate the possibility of any undue importance given to some features/variables over others.

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

   Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variable, VIF for a regression model variable is the ratio of overall model variance of a model that involves that a single independent variable.
   Essentially VIF becomes infinity when the denominator tends towards 0 or the independent variance of a model containing a single independent variable yield towards 0 indicating perfect correlation.
   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).

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

   A Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

   Quantiles are points in a distribution that relate to the rank order of values in that distribution.

   For a sample, you can find any quantile by sorting the sample. The middle value of the sorted sample (middle quantile, 50th percentile) is known as the median. The limits are the minimum and maximum values. Any other locations between these points can be described in terms of centiles/percentiles.