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?

The categorical variables available in the dataset are season, mnth (month),yr (year), holiday, weekday, workingday, weathersit. Based on the boxplot/barplot analysis of these variables with respect to the target variable cnt we can conclude the following

- season: More bookings have happend in the Fall season. We see increase in the average bookings from the seasons of spring to fall. After the fall we see a dip in the average bookings. We also observe year over year increase in bookings across all seasons
- yr: There is an increase in the business from 2018 (0) to 2019 (1).
- mnth: We see an increase in average bookings from Jan till June before it starts to decrease in 2018. We see a similar pattern in 2019 with the rise seen till September. One anomaly is we see a dip in July when compared to June in 2019
- holiday: We see less bookings on average on holidays compared to when its not an holiday
- weekday: Bookings on avegrage gradually increses from the start of the week and start to dip towards the end. During 2018 however, there is not much variation in the bookings during the week
- workingday: Bookings on working day are more on average. This maybe an indicator that people take bikes to work. This correlates with why we see less bookings during holidays
- weathersit: As expected on a clear sky we have more bookings on average. We have least bookings on average during light rain/snow. There are no bookings on heavy rain/heavy snow as expected

Based on the observations we can clearly see there is a liner relationship between the categorical variables and our target cnt variable and we can therefore model it as a linear regression model.

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

Dummy variable creation is a process to convert a categorical variable into 0/1 value so that it can be fed into the linear regression model. A categorical variable with K unique values can be represented using K-1 variables.

If drop_first is not set to true the following method will create K variables. One extra which is actually doesn't serve any purpose and is just extra.
`pd.get_dummies(df['cat_column'])`

Hence, we use `pd.get_dummies(df['cat_column'], drop_first=True)` to remove the extra column

Lets take an example of season to explain this. If we don't `drop_first`, we get an output like this

spring	summer	fall	winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Here for a row, the column which has 1 is the season. So, for the first row the season is spring, summar for the 2nd row, fall for the 3rd row and winter for the 4th row

But the same can be explained by one less column as follows with `drop_first=True`

summer	fall	winter
0	0	0
1	0	0
0	1	0
0	0	1

The same logic applies here. The column with 1 is the season. Since the 1st row has no 1 it automatically implies its spring

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

The variable `temp` has the strongest positive correlation with the target variable `cnt`

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

Following assumptions of a linear regression model were validated

- **There is a linear relationship between X (features) and Y (target variable)**: This was validated by plotting the scatterplot of the feature columns with the target variable `cnt`
- **Error terms are normally distributed with mean zero**: This was validated by plotting the histplot of y_train - y_train_pred. The destribution was normal and centered around 0
- **Error terms are independent of each other**: This was validated by plotting the scatterplot of error terms with the target variable. No visible pattern

- **Error terms have constant variance (homoscedasticity)**: This can be validated as variance in the error terms remains constant using the scatterplot of error terms and target variable
- **Multicoliearity validation**: Of the variables selected in model building, we have validated the VFI score for each variable is below 5

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

Based on the absolute values of the coefficients of the features,we can say the top 3 features contributing significantly to the demand are

1. temp
2. light_rain_snow
3. yr

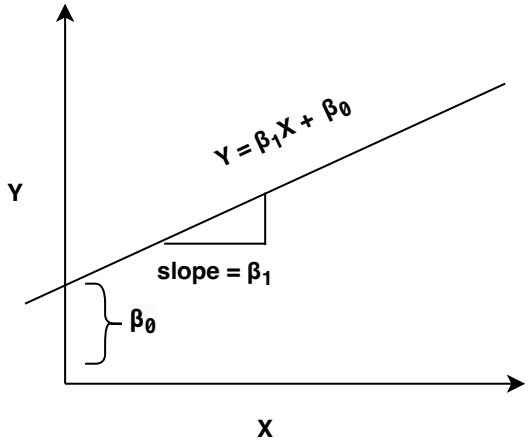
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model establishing a linear relationship between a dependent variable called the **Target Variable** and one or more independent variables called **Feature**. The linear relationship is obtained by fitting a line equation through the data. The objective is to find the best fitting line (regression line) such that it minimizes the sum of squared differences between the observed values and the values of the **Target Variable**

The equation of the linear regression line is as follows (Single linear regression)

$Y = \beta_1 X + \beta_0$



Here β_0 denotes the intercept of the line and β_1 is the slope. A positive β_1 means Y (target variable) increases with increase in X and vice-versa.

This can be generalized to multiple linear regression (multiple independent variables)

$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \beta_0$

The coefficients are computed in such a way that we minimize on the sum of squres or error terms

Some important assumption have to be made about the data by liner regression

- There is a linear relationship between X & Y
- Error terms ($Y - Y_{pred}$) are normally distributed with zero mean
- Error terms are independent of each other. There is no pattern to them
- Homoscedasticity of error terms i.e constant variance of error terms
- It is assumed that there is very little or no multicollinearity between the predictor variables - i.e they are independent

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet refers to four datasets with identical or near identical statistics properties like mean, variance, number of data points, correlation between independent and target variable and also the same regression line equation yet have very different distributions. This was created by statistician Francis Anscombe to ephasize the importance of visualization of the data and how their distributions and outliers can have affect on the statistical properties

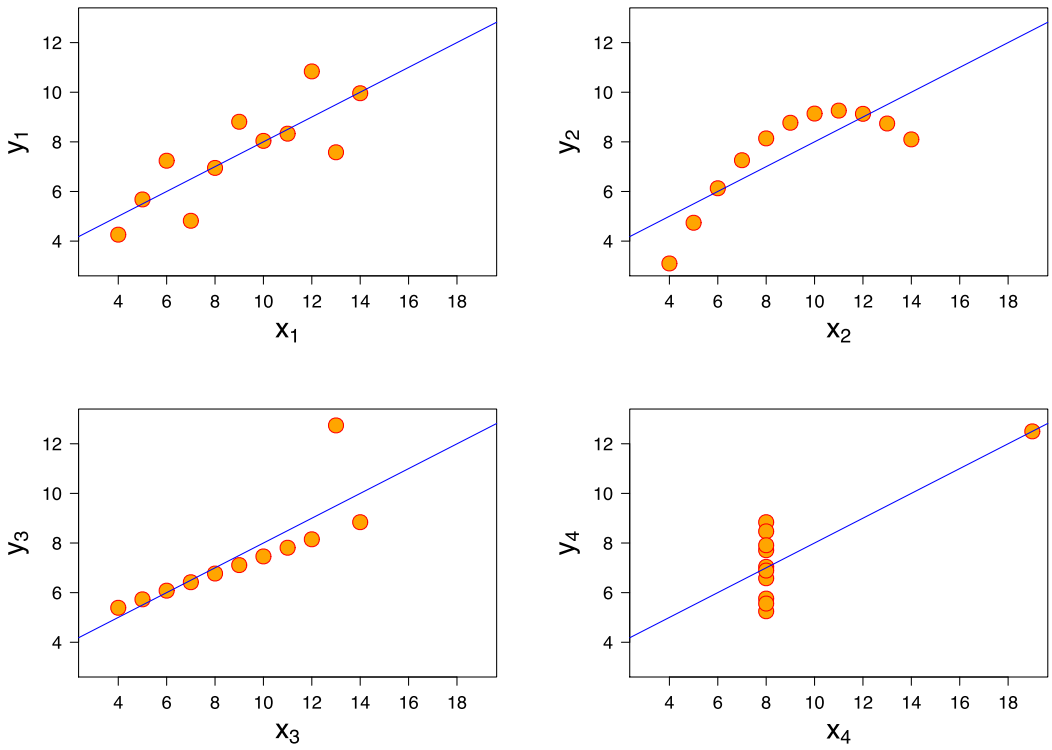
Example from Wikipedia of the quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Statistical properties analysis for the quartet

Property	Value	Accuracy
Mean X	9	Exact
Mean Y	11	Exact
Variance X	7.5	to 2 decimals
Variance Y	4.125	±0.003
Correlation between X & Y	0.816	to 3 decimal places
Linear Regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
R2	0.67	to 2 decimal places

But if we observe, there distribution are completely different

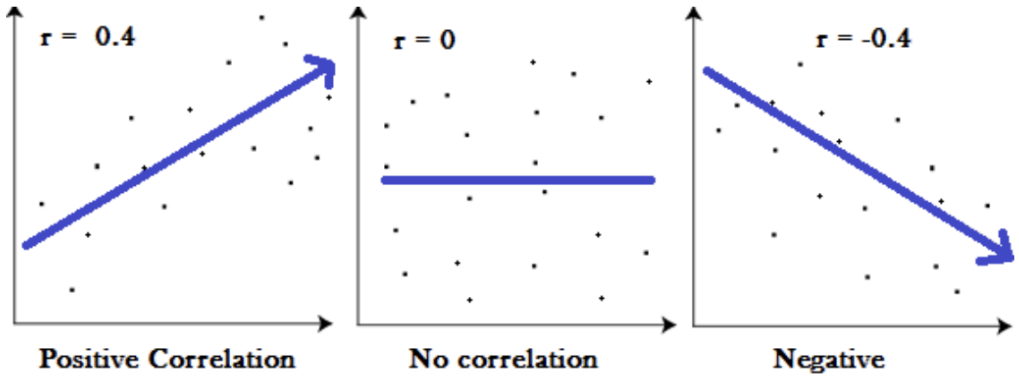


- [Top Left] Dataset one has a linear relationship
- [Top Right] Clearly Dataset two is not linear
- [Bottom Left] Dataset three's outlier has had an affect on its regression line
- [Bottom Right] Dataset four's outlier seem to have establised a linear relationship which shouldn't be there. (High leverage point outlier)

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (r) is a correlation coefficient that measures linear correlation between two sets of data. The value lies in [-1,1]. A very high value means that there is positive correlation between the datasets i.e increase in one results in increase in the other and vice-versa. Similarly very negative value implies negative correlation i.e increase in one results in decrease of the other and

vice-versa. A value of 0 indicates no correlation i.e both the datasets are indepenent of each other and one doesn't influence the other.



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

Feature scaling is a pre-processing technique of bringing the dataset variables within a fixed range. A dataset can have lot of features with units and values. As a result the linear regression may have coefficient that are not comparable and features with higher values may just outright dominate other fetures with lower values. Feature scaling by bringing the features into a fixed range helps mitigate this problem.

A simple example to demostrate why feature scaling is important. Lets say we have a dataset to predict the cost of house (in Rs). One of the variable is area in (sqft). Here datapoints can be Price=10000000 , area=2000

We can see how the scales of both the variables are very different. This difference will definitely affect the coefficient values of the regression line.

There are many ways to scale the data. The most common ones are normalized scaling & standardized scaling

Normalized Scaling	**Standardized Scaling**
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1	Scales the values to have a mean (μ) of 0 and a standard deviation (σ) of 1
Range [0,1]	No defined range
Suceptible to presence of outliers	Not Suceptible to outliers
Keeps the original distribution of data	Modifies the distribution of data. Makes it a normal distribution
Suitable when data is within a range with no extreme outliers	Suitable when the distribution is normal and if outliers are present
Formula: $X_{normalized} = (X - X_{min})/(X_{max} - X_{min})$	Formula: $X_{standardized} = (X - \mu) / \sigma$

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

The Variance Inflation Factor (VIF) for a predictor variable is a measure of how correlated it is to the other predictor variables i.e how well it can be explained by other predictor variables

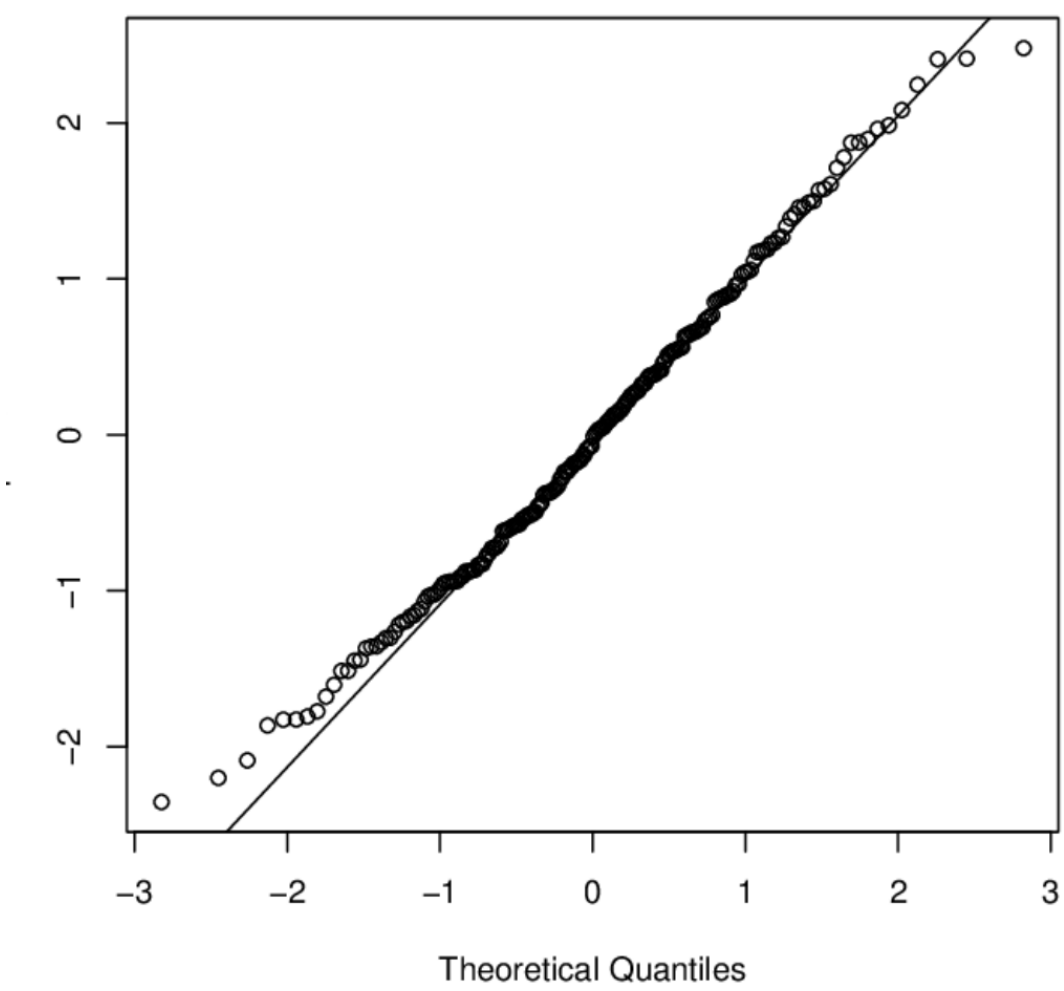
$VIF_i = 1 / (1 - R^2_i)$

In an scenario when the variable i is perfectly correlated to the other predictor variables, then $R^2_i = 1$ and hence VIF_i will be infinity

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

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a specific distribution like a normal distribution. It compares the quantiles of the observed data with the quantiles expected from the theoretical distribution. We plot a reference line with slope 1 (45^o). Then we plot the quantiles of the dataset. If the points fall approximately on the line then we conclude the distribution of the dataset match with the normal distribution. If there is a high diversgence from that line we conclude the distribution is not a normal distribution.

The Q-Q plot is particularly useful for checking the normality assumption in linear regression. It is also useful to check if residuals have a constant variance, which is an assumption for the homoscedasticity of the model


Normal Q–Q Plot
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles