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?

There are several categorical variables in the dataset provided.

- **Year**: The median count of people who used the service increased in 2019 compared to 2018. This shows that the demand is on the rise
- **Season**: People prefer to use the bike more in fall and summer compared to spring and winter.
- **Holiday**: Demand for shared bikes is higher on a working day. This shows that people are using the bike service for travelling to work
- **Weekday and Workingday**: convey almost the same information. That the demand for bikes is very similar on all days
- **Weathersit**: Demand for bikes is highest on a clear day, followed by a misty+cloudy day. Least demand is on days when it snows/rains lightly.
- **Mnth**: This variable indirectly correlates with season. The demand is highest during the months of may, june, july, august and september

When RFE is used to shortlist the columns, most of these columns have high coefficient, low p-value, low VIF value and the rankings provided show that these columns have a significant effect on the target variable

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

In order to reduce the number of predictor variables. If there are n categories in the column, we only need n-1 number of columns to represent them.

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

“temp” and “atemp” variables have the highest correlation with target variable.

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

- **Linear relationship between predictor variable and target variable.**
- **Minimal multi-co-linearity between variables. Pairplot shows this information.**
- **Homoscedasticity Assumption: There shouldn’t be any clear pattern between residual values and predicted values.**
- **Normal distribution of error(residual) terms**

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

When the RFE was used to select/prune the number of features, below 3 variables had the highest coefficient values:

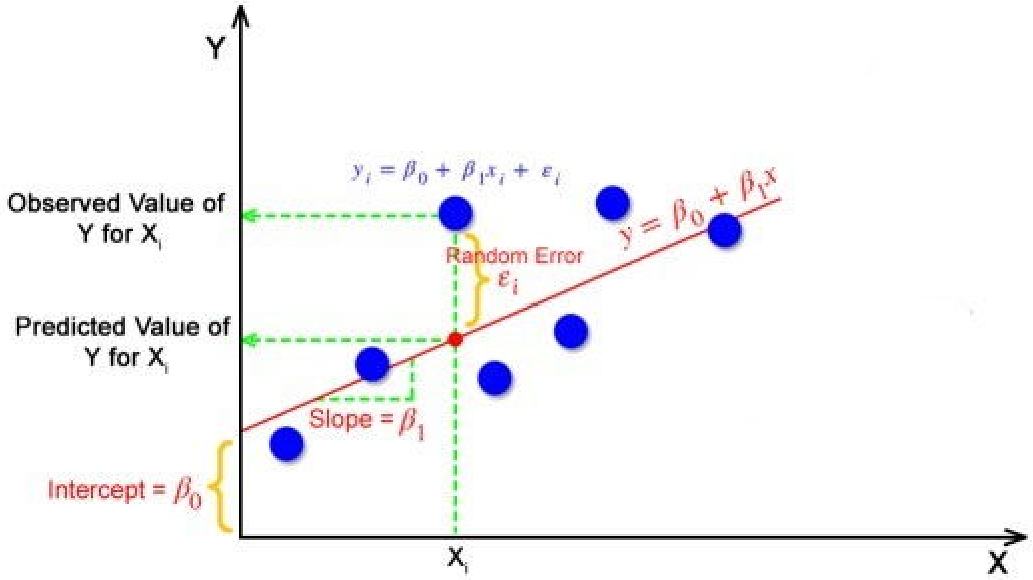
- yr
- atemp
- winter

General Subjective Questions:

1.**Explain Linear Regression in detail**

Regression analysis is a statistical method to determine the relationship of 2 variables. Variable we want to predict is called Dependent variable and the variable used for prediction is called Independent/predictor variable. There can be more than one independent variable used for analysis. In this case it is called Multiple Linear Regression. Linear regression assumes a linear relationship between the target and predictor variables and tries to find the best fit line. This line is derived using the Ordinary Least Squares method. OLS is the absolute difference between the predicted y value and actual y value which is then summed.

y = b0 + b1x1 + b2x2 + …..
b0 -> intercept
b1 -> Coefficient of variable x1



Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of Y for X_i
Random Error ε_i
$y = \beta_0 + \beta_1 x$
Predicted Value of Y for X_i
Slope = β_1
Intercept = β_0
X_i
X

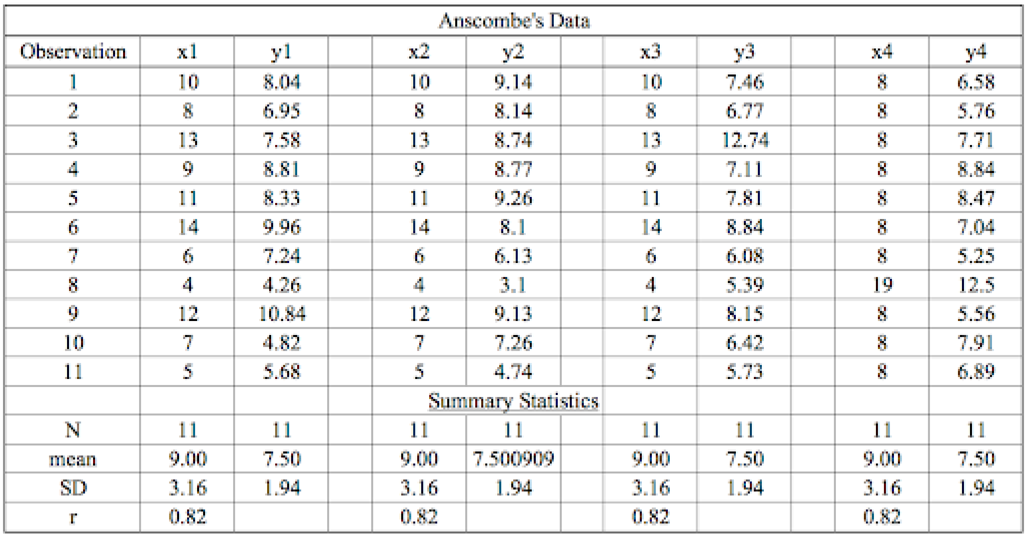
Gradient descent is one method used to optimize the cost function to reach the optimum minimal solution.
R-Squared(R2) is used as an evaluation metric to analyse the performance. It explains the amount of variation that is captured by the model. Ranges between 0 & 1. Higher the value, the better the model fits the data.

2.Explain the Anscombe’s quartet in detail?

Anscombe’s Quartet is a group of four datasets which are nearly identical in simple descriptive statistics but there are some peculiarities that fools the Regression model if built. They have very different distributions and appear differently when plotted on scatter plots. These 4 dataset will have similar stattiscts such as:

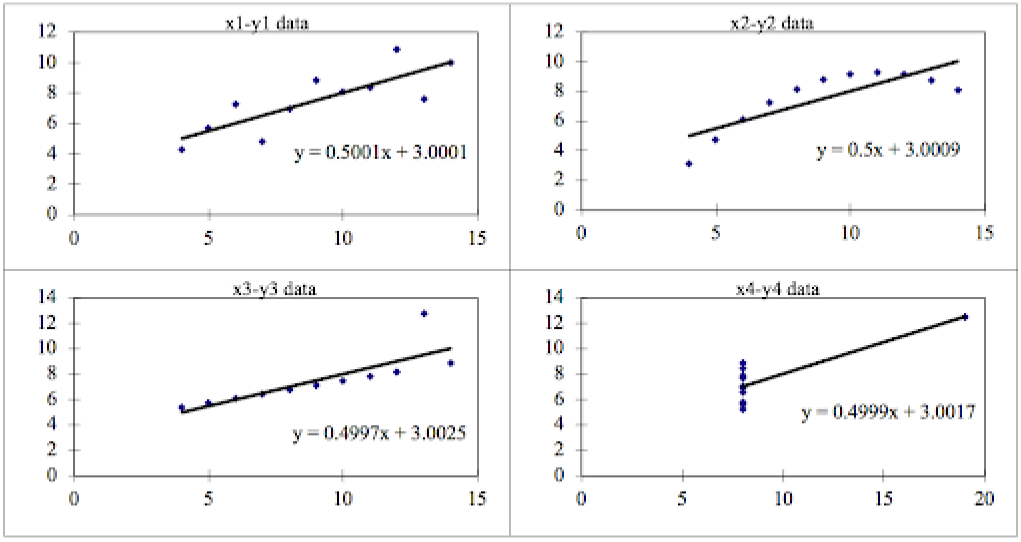
- Number of samples
- Mean
- SD
- R

However, when these models are plotted on a scatter plot, each dataset generates a different kind of plot that isn’t interpretable by a regression algorithm.



Anscombe's Data

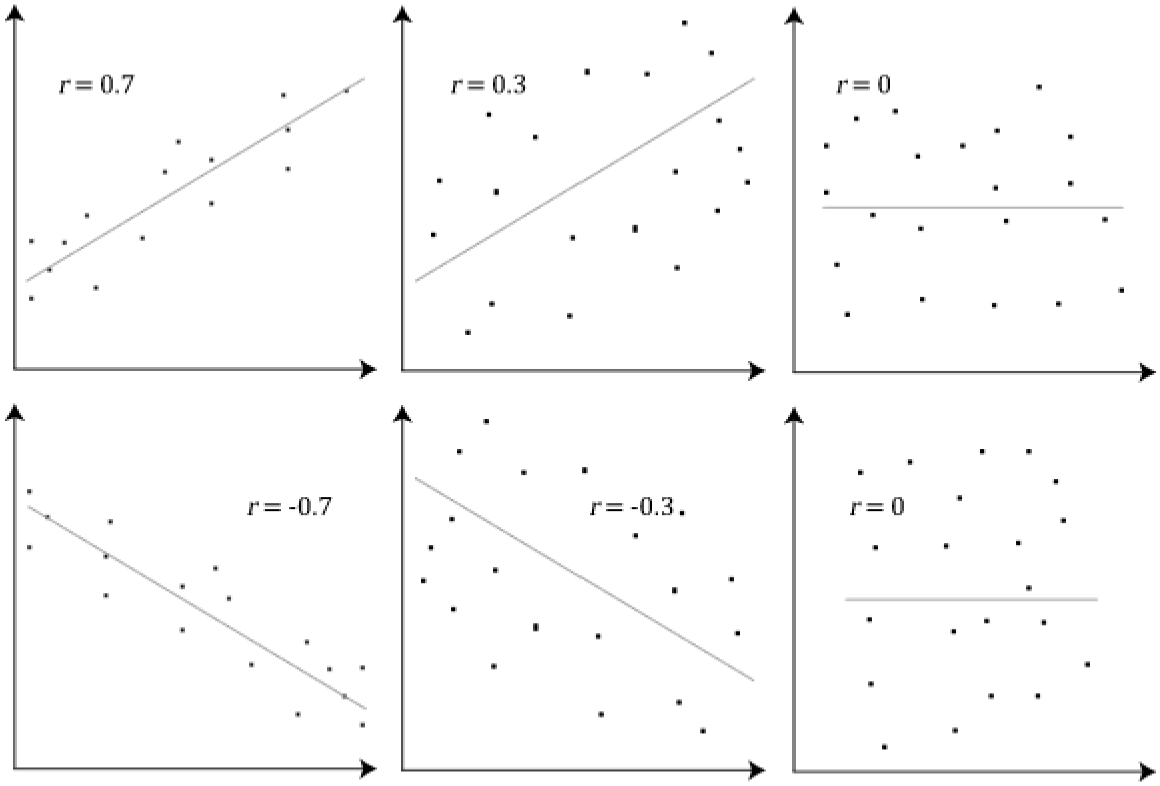
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	



Anscombe’s quartet helps us understand the importance of data visualisation and how easy it is to fool a regression algorithm.

3. What is Pearson’s R?

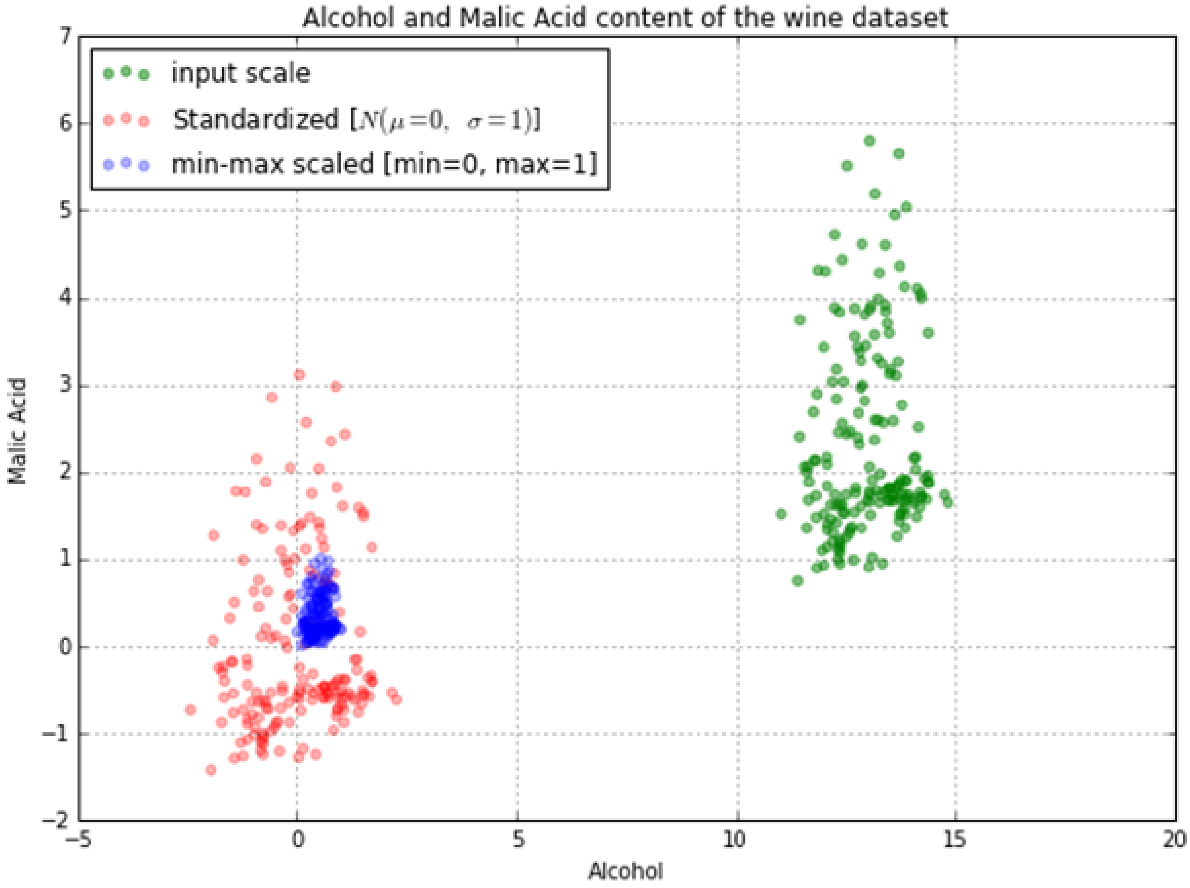
It is a statistic that measures the linear correlation between 2 variables. The value of R ranges between -1 and +1.
A value of 0 indicates that there is no correlation between the variables
A value > 0 indicates positive correlation between the variables
A value < 0 indicates negative correlation.



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

The goal of applying feature scaling is to make sure features are on almost the same scale so that each feature is equally important and make it easier to process by most machine-learning algorithms

In Standardised scaling, the features will be rescaled to ensure that the mean is 0 and the standard deviation is 1.
In Min-Max scaling or Normalised scaling, the features are rescaled with a distribution value between 0 and 1.



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

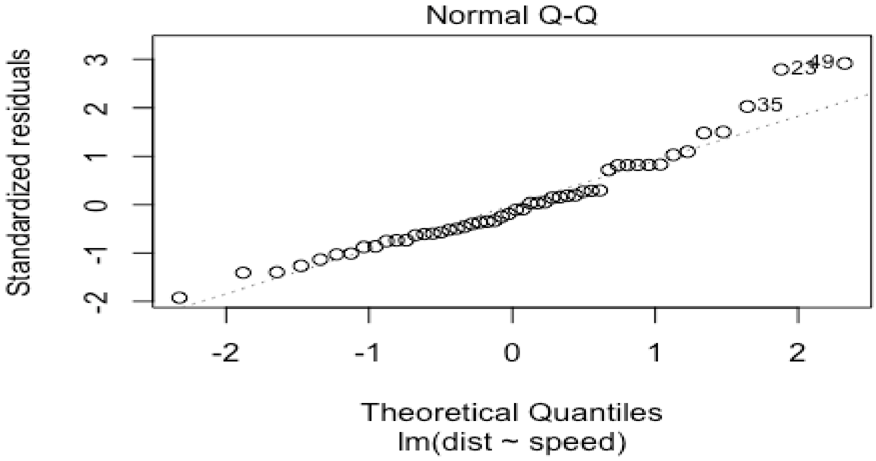
Variance Inflation Factor measures the severity of multicollinearity in the OLS regression method.
If all the independent variables are orthogonal to each other, then VIF = 1.0.
If there is perfect correlation then VIF = inf.
VIF = 1 → No multicollinearity
VIF < 5 → Moderate correlation
VIF >10 → Severe

If VIF is large, one approach is to review the independent variables and remove those that are duplicates or are not adding value to explain the variance in the model.

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

Quantile-Quantile 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 us determine if 2 datasets came 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 the same distributions.


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(dist ~ speed)