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)

A: The analysis is done on categorical variables using Count and Box plots. We can infer below point from those visualizations

- Most of the bookings has been done during the month of May, June, July, Aug, Sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year
- When it's not holiday, demand seems to be less in number which seems reasonable due to most of the customers spending their time with family and friends
- Thursday, Friday, Saturday and Sunday have more bookings compared to other days
- Demand for bikes is more on non - working day
- When the weather sit is Clear the demand for bikes is more
- Season 'Fall' have more demand in bikes booking when compare to other Seasons
- The demand for bikes increased in 2019 when compared with previous year 2018

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

A: Basically drop_first=True will help in reducing the extra column created during dummy variable creation which interim helps in reducing the correlation created among dummy variables.

Syntax: drop_first= bool, default False

Whether to get k-1 dummies out of k categorical levels by removing the first level.

For e.g.: In our use case for categorical column 'Season' we created the dummy variable because we want to get the categorical column converted into numeric column which helps in including this column in our analysis. Once we create the dummy variable there is no need to have the 1st Unique column which is 'Season' in our dataset because of that we use drop_first=True which will drop this categorical column from dataset.

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

A: 'Temperature' variable is having highest correlation with target variable.

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

A: Whenever we perform Linear Regression the basic assumptions we make are below:

- Normal Distribution of error terms
- Multicollinearity check: Here we will validate is there any multicollinearity among the variables
- Linear relationship validation which means validating of Linearity among the variables
- Homoscedasticity: Here we validate is there any visual patterns in residual values
- Independence of Residuals: No auto correlation

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

A: From final model I found that below 3 features are significantly contributing towards the demand of shared bikes among other features

- Temperature
- Winter
- Month – September

General Subjective Questions

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

A: Linear Regression algorithm is a statical model that analyse the linear relationship between Dependent and In-Dependent variable which helps in predicting the future outcomes of the given data which means whenever there is a change in any of the independent variables the dependent variables will also change accordingly it can be a decrease or increase.

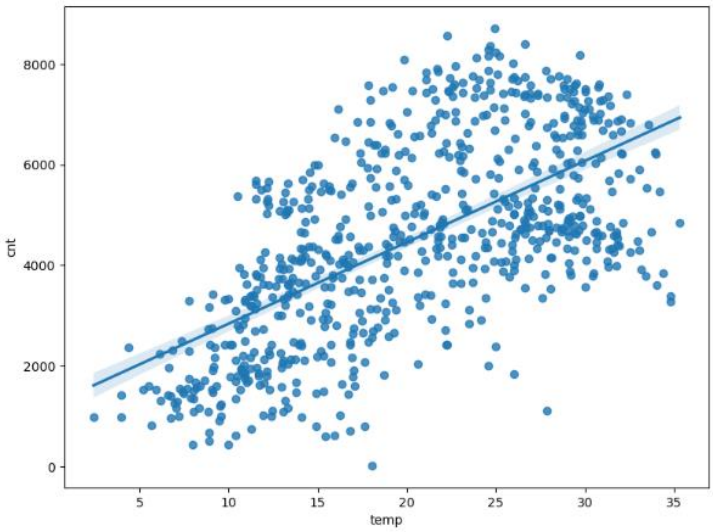
The equation to calculate Linear regression is:

y = c+mx

where y is the dependent variable to predict, c is the intercept, m is coefficient of that feature which is slope of the regression line and x is the independent variable we use to predict.

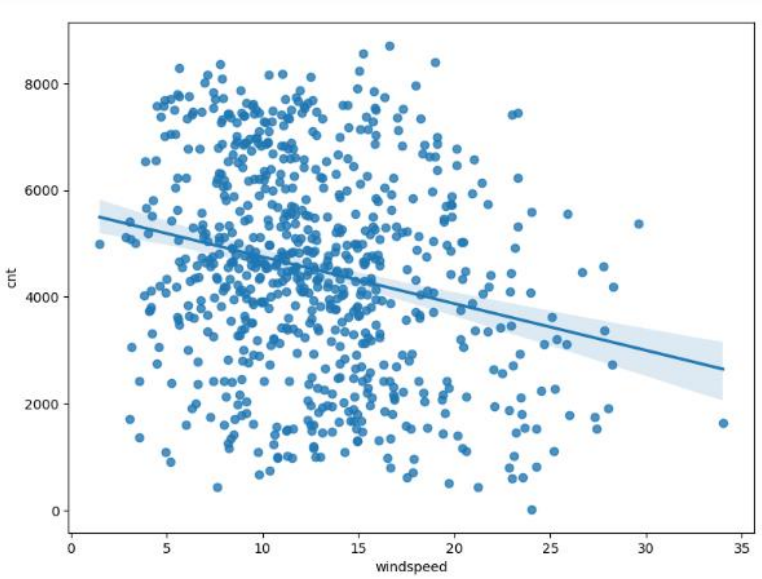
Linear regression can be both Positive and Negative in nature.

- Positive Linear Regression: If the linear relationship between the dependent and independent variables increases if any one increase, then it is called Positive Linear Regression.
  E.g.:



In the above Regplot we can observe the value of 'cnt' variable increase with 'temp' variables which shows positive linearity.

- Negative Linear Regression: This is opposite to Positive Linear regression where the linear relationship between the dependent and independent variables decreases if one increase or vice versa.
  E.g.:



  In the above Regplot we can observe the value of 'cnt' variable decreased when the 'windspeed' variable increased which shows negative linearity

Linear Regression is of 2 types:

- Simple Linear Regression: In this we will perform analysis between one dependent and independent variable
- Multiple Linear Regression: In this we will perform analysis between one dependent variable and multiple independent variables.

While performing Linear Regression we make the following assumptions:

- Normal Distribution of error terms
- Multicollinearity check: Here we will validate is there any multicollinearity among the variables
- Linear relationship validation which means validating of Linearity among the variables
- Homoscedasticity: Here we validate is there any visual patterns in residual values
- Independence of Residuals: No auto correlation

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

A: Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

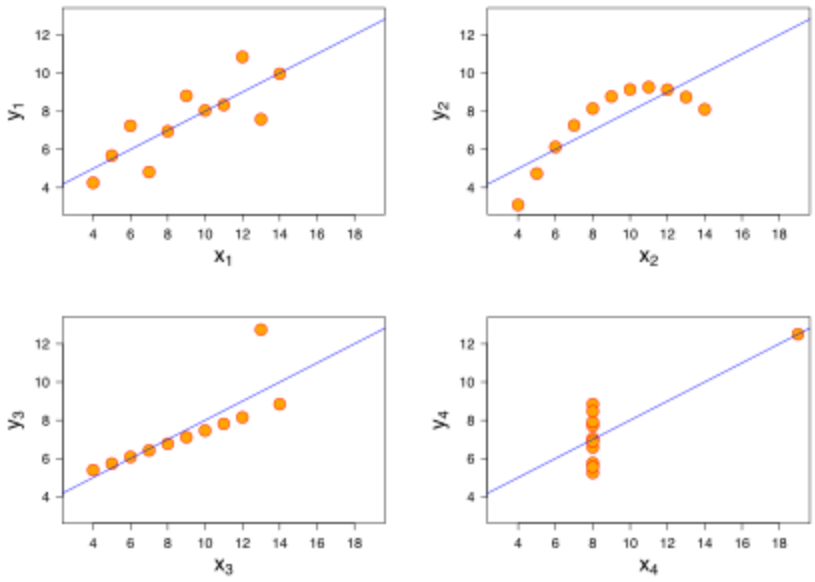
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset. When we plot these four datasets on an x/y coordinate plane, we can notice that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

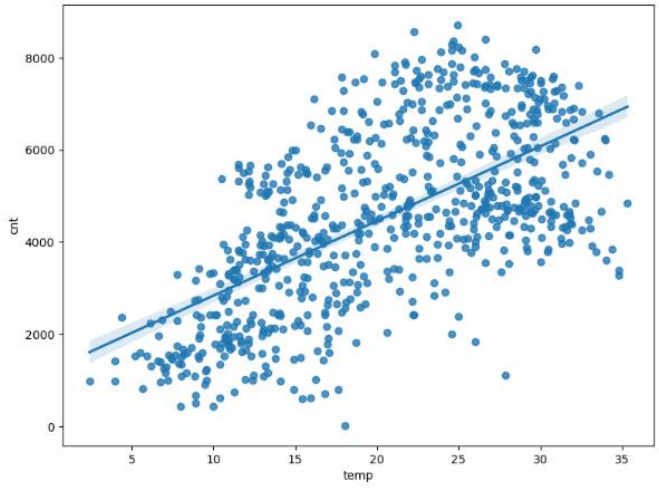
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

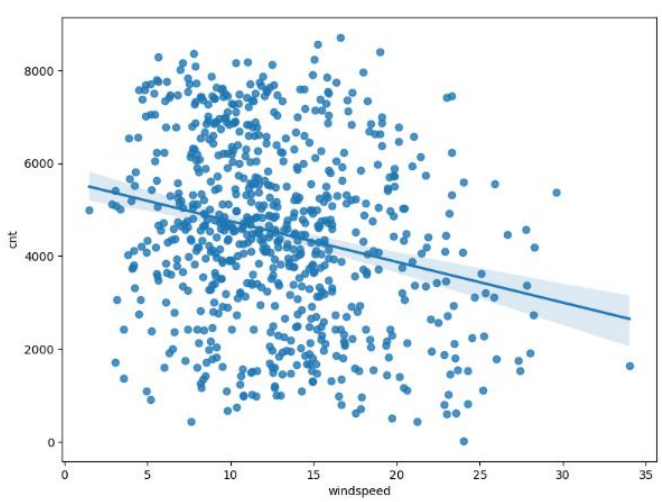
A: Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to increase and decrease together, the correlation coefficient will be positive. If the variables tend to increase and decrease in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, R, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:

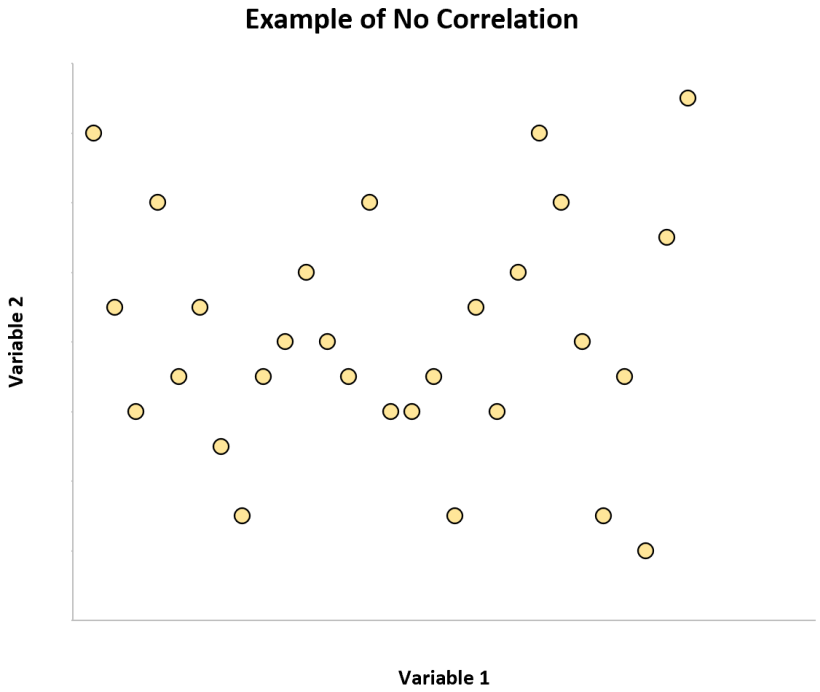
Positive:



Negative:



No Correlation:



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

A: Scaling is the technique that helps in standardizing the independent variables/features present in dataset for a fixed range. Scaling doesn't impact the model. In other words Scaling is a technique which helps in re-scaling feature values with the distribution value between 0 and 1 which is useful for the optimization algorithm such as Gradient Descent. Scaling is performed during data pre-processing step to handle varying values or units highly. If we don't perform scaling then we need up in getting some of the coefficients obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So, it is advised to use scaling so that the units of the coefficients obtained are all on the same scale.

Scaling can be performed in 2 ways:

- Normalization (Min-Max Scaling)
- Standardization (mean-0, sigma-1)

Difference between Normalized and Standardized Scaling

Normalized Scaling	Standardization Scaling
Min-Max values of the features are used for scaling and Scales between -1 to 1	Mean and Sigma values of the features are used for scaling and this is Not bounded to certain range
Applied when the features having separate scales	Applied when we verify the Mean is 0 and unit Sigma (Standard Deviation)
Highly affected by Outliers	Impact of Outliers is less

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.
This is often called Scaling Normalization	This is often called Z-Score Normalization
When the feature distribution is not clear then this will help	When we have the clear feature distribution then this will help

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

A: If there is perfect correlation, then VIF = infinity. 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 multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independents variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

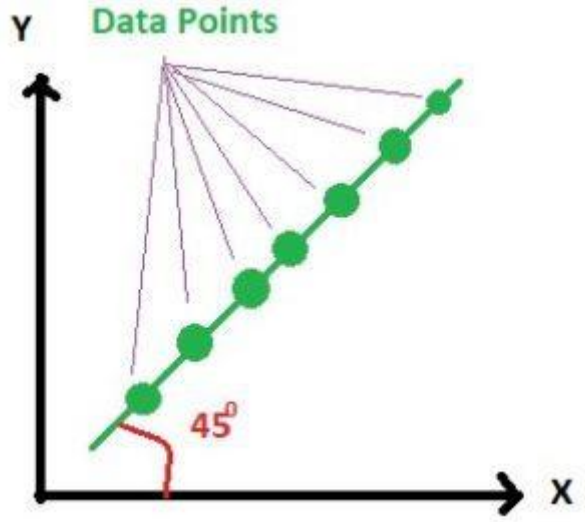
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot. This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Sample Q-Q Plot:



Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.