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?
   Answer::

   We have categorical and numerical independent variables. I have used different types of plots such as bar, box, pair to identify the effect on the dependent variables. Major insights are mentioned below.
   - Number of booking has been drastically increased in 2019 in comparison with 2018
   - Most booking happens on Fall season
   - Booking trend increase from starting till mid year and then it started decreasing. This is the same trend in both years
   - Demand is less on Holidays and is understandable
   - Demand is relatively more on days approaching weekends
   - Clear weather is having more booking in both years
   - booking seems to be same on working day and non-working day
   - There is a positive correlation between count and temp, count and atemp. Weak negative between count and humidity, count and windspeed.
   - Demand for bike is positively correlated with temperature
   - Count is decreasing with increase in humidity
   - Count is decreasing with increase in windspeed
2. Why is it important to use drop_first=True during dummy variable creation?
   Answer ::

   In general we have to follow Occam's Razor problem solving principle as it is essentially stating to pick the simple solution for any problem. One of the approach to keep our model simple is to have less number of variables. At the same time we need to convert categorical variables to dummy variables as computers understand only numbers.
   drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

   Drop_first keeps n-1 predictors for n different values of a categorical predictor.

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

   Answer::

   Temp variable has the highest correlation.

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2080      0.030      6.829      0.000       0.148       0.268
yr             0.2349      0.008     28.418      0.000       0.219       0.251
holiday       -0.0956      0.026     -3.656      0.000      -0.147      -0.044
temp           0.4583      0.034     13.435      0.000       0.391       0.525
windspeed     -0.1555      0.025     -6.119      0.000      -0.205      -0.106
season_spring -0.0500      0.021     -2.391      0.017      -0.091      -0.009
season_summer  0.0583      0.014      4.070      0.000       0.030       0.086
season_winter  0.0876      0.017      5.157      0.000       0.054       0.121
mnth_jan      -0.0403      0.018     -2.242      0.025      -0.076      -0.005
mnth_sep       0.0900      0.016      5.527      0.000       0.058       0.122
weathersit_Light_rain -0.2863      0.025    -11.538      0.000      -0.335      -0.238
weathersit_Misty -0.0779      0.009     -8.864      0.000      -0.095      -0.061
==============================================================================
```

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

   Answer::

   Following assumptions have been validated after building model.

   - Normality of error terms – Error terms are normally distributed
   - Multicollinearity check – there are no significant relationship between independent variables
   - Linear relationship – There are linear relationship between independent and dependent variables
   - Homoscedasticity – Residual variance is equally distributed.

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

   Main 3 features are ::

   Temp – Temperature is having a positive correlation
   weathersit_Light_rain – Light Rain weather is having a negative correlation
   yr – Year is having a positive correlation.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2034	0.030	6.877	0.000	0.145	0.261
yr	0.2339	0.008	28.414	0.000	0.218	0.250
temp	0.4917	0.033	14.805	0.000	0.426	0.557
windspeed	-0.1497	0.025	-5.963	0.000	-0.199	-0.100
mnth_July	-0.0483	0.019	-2.587	0.010	-0.085	-0.012
mnth_September	0.0723	0.017	4.260	0.000	0.039	0.106
season_spring	-0.0682	0.021	-3.227	0.001	-0.110	-0.027
season_summer	0.0479	0.015	3.145	0.002	0.018	0.078
season_winter	0.0818	0.017	4.739	0.000	0.048	0.116
weekday_sunday	-0.0450	0.012	-3.847	0.000	-0.068	-0.022
weathersit_light	-0.2847	0.025	-11.513	0.000	-0.333	-0.236
weathersit_mist	-0.0802	0.009	-9.161	0.000	-0.097	-0.063

General Subjective Questions

1. Explain the linear regression algorithm in detail?

   Answer::

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

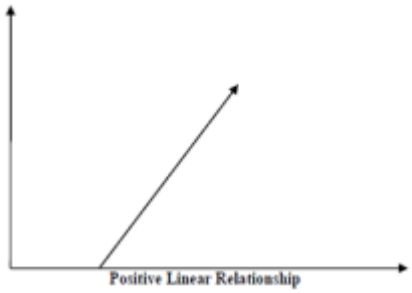
a = y-intercept of the line

x = Independent variable from dataset

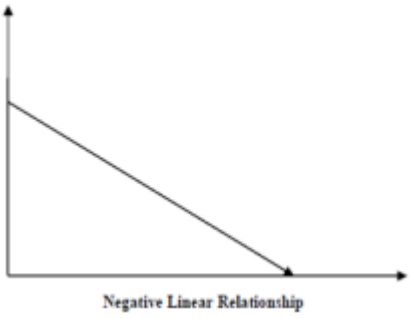
y = Dependent variable from dataset

Furthermore, the linear relationship can be positive or negative in nature as explained below:

- Positive Linear Relationship:
  A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph



- Negative Linear relationship:
  A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph.



Linear regression is of the following two types –

Simple Linear Regression

Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

Multi-collinearity –

Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation –

Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables –

Linear regression model assumes that the relationship between response and feature variables must be linear.

Normality of error terms –

Error terms should be normally distributed

Homoscedasticity –

There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail?

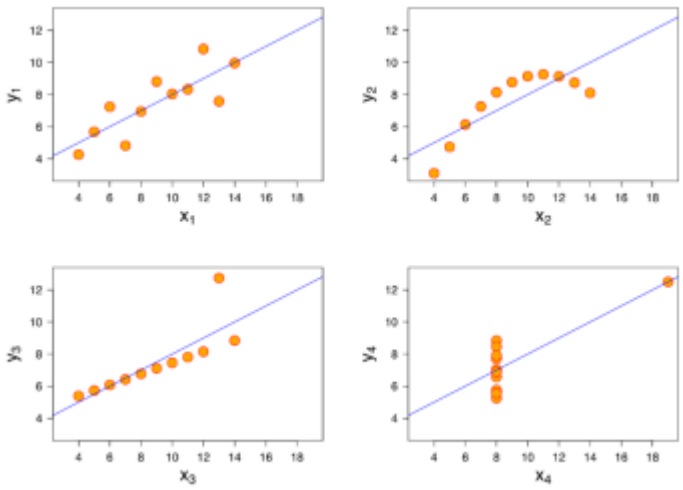
Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician *Francis Anscombe* in 1973 to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis**.**

	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

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 observe that they show
the same regression lines as well but each dataset is telling a different story:

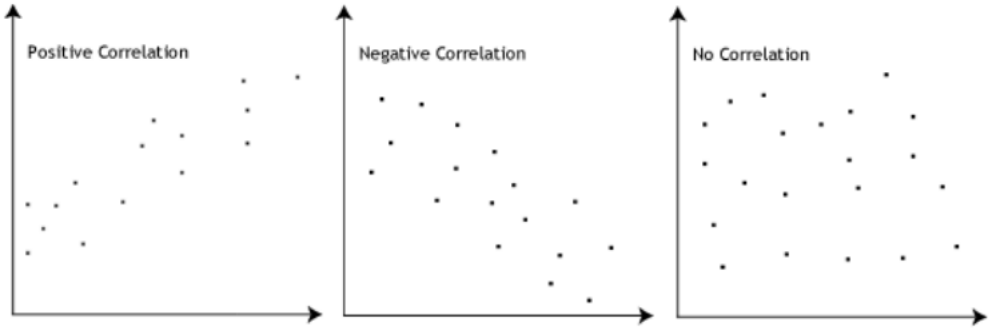


• Dataset I appears to have clean and well-fitting linear models.
• Dataset II is not distributed normally.
• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.
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?

Pearson's R is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

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:



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

   Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

S.NO.	Normalization	Standardization
6.	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.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

   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 independent 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 ?

   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

   Few advantages:
   a) It can be used with sample sizes also

   b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

   It is used to check following scenarios:

   If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behaviour

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.