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)

   Ans. Some of the inferences that can be made from the analysis of the categorical variables from dataset on the dependent variable that is Count-

   i. Bookings are nearly the same for working and non-working day even though the spread is bigger for non-working days people don't rent out bikes as people might have some plans.

   ii. Fall is better for riding bikes as expected in the weather condition that is optimal. The fall has the highest median, henceforth, there are more booking in fall.

   iii. More bookings are there in clear weather as temperature and humidity is optimal.

   iv. There are more number of bookings in Thursday, Friday, Saturday and Sunday as compared to other days.

   v. People like to spend time with family and may use their own personal transportations instead of bikes, so people rent more on non-holidays compared to holidays.

   vi. As in year 2019 year has higher median than 2018's bike rents and increasing yearly, this can be because of increasing popularity and awareness of bike rentals.

   vii. May, June, July, Aug, Sept and Oct had most of the bookings. It increased till mid of the year and then decreased gradually.

   viii. Season plot is similar to the month plot's spread as fall has higher median.

   ix. Saturday and Wednesday has bigger spread that means, those who have plans for Saturday might not rent bikes as it a non-working day. Overall median across the week is same.

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

   - Drop first is True is used to reduce any extra column created during dummy variable creation and to remove the redundant data. It also helps to avoid multi-collinearity created among dummy variables.
   - Drop first is False implies that it gets k-1 dummies out of k categorical levels by removing the first level. If there are three types of Categorical column's values, then there is a need to create the columns'' dummy variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - Temperature (Temp) value has significantly high correlation with Count (Target Variable).
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Five assumptions of Linear Regression Model are validated by –

- Normality of error terms - Normal distribution should be followed by residual error terms.
- Multicollinearity check - An Insignificant multicollinearity among variables should be present.
- Linear relationship validation – Linear relationship between the dependent variable (Test and Predicted). Linearity among variables visibility should be present.
- Homoscedasticity - Residual values shouldn't have a visible pattern
- Independence of residuals - No auto-correlation between the residuals

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

   Top three features that contribute significantly towards explaining the demand of the shared bikes –

   - Temperature(0.4354)
   - Weather Situation-Light and Snowy(0.2837) i.e. winter
   - Year(0.2461) – for september

General Subjective Questions

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

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. It that analyses the linear relationship between a dependent variable with given set of independent variables. Regression models a target prediction value based on independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease). It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

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 the relationship can be represented with the help of following equation −

$Y = mX + c$

Here, Y is the dependent variable we are trying to predict.

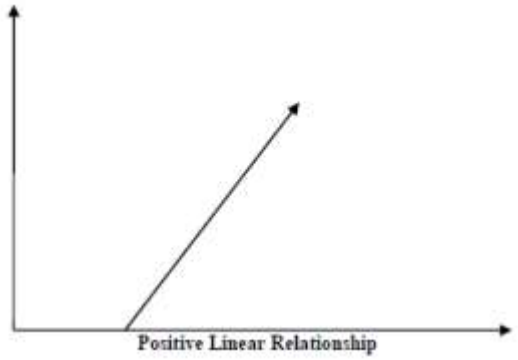
X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

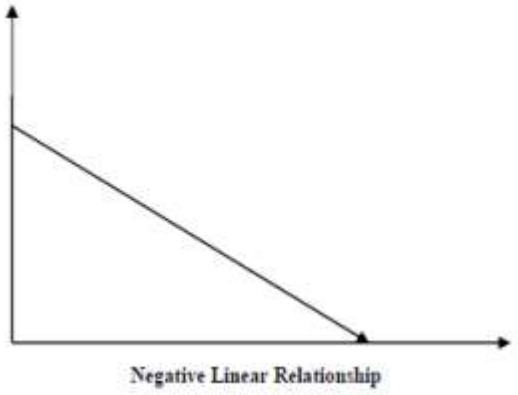
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

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

An example is let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Y=theta1 + theta2*x

While training the model we are given :
input training data (univariate – one input variable(parameter))
labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best
θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

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. (3 marks)

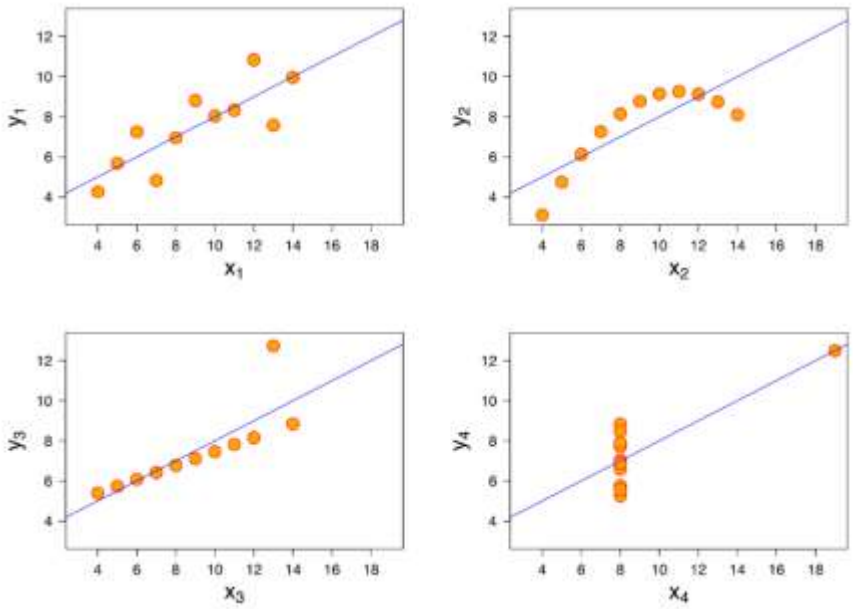
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed. One of these, the Datasaurus Dozen, consists of points tracing out the outline of a dinosaur, plus twelve other data sets that have the same summary statistics. Datasaurus Dozen was created by Justin Matejka and George Fitzmaurice. The process is described in their paper "Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing". The Datasaurus Dozen proves us as much as Anscombe Quartet why visualizing our data is important as summary statistics can be the same, while data distributions can be very different. 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	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.

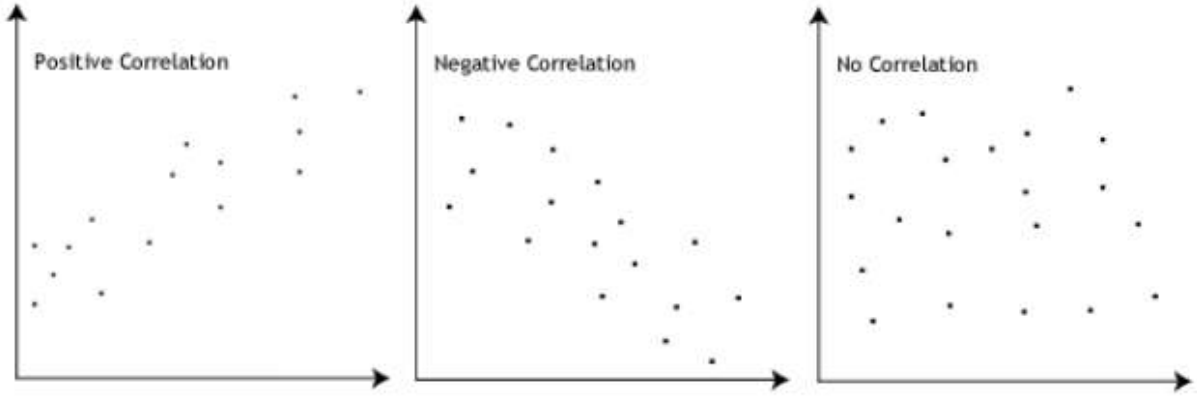
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

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

In statistics, the Pearson correlation coefficient — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — 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 relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation). 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:



Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

For a population

Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables {\displaystyle (X,Y)}, the formula for ρ

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- is the covariance
- is the standard deviation of
- is the standard deviation of

The formula for rho can be expressed in terms of mean and expectation. Since

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

the formula for rho can also be written as

$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Some of the practical issues are - extracting the correlation coefficient between two sets of stochastic variables is nontrivial, in particular where Canonical Correlation Analysis reports degraded correlation values due to the heavy noise contributions under heavy noise conditions. A generalization of the approach is given elsewhere. In case of missing data, Garren derived the maximum likelihood estimator.

Some of the mathematical properties - The absolute values of both the sample and population Pearson correlation coefficients are on or between −1 and 1. Correlations equal to +1 or −1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation).
The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X). A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may transform X to a + bX and transform Y to c + dY, where a, b, c, and d are constants with b, d > 0, without changing the correlation coefficient. (This holds for both the population and sample Pearson correlation coefficients.) Note that more general linear transformations do change the correlation: see § Decorrelation of n random variables for an application of this.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. . 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. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).sklearn.preprocessing.scale helps to implement standardization in python. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

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

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. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/ (1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. 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). 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. (3 marks)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Q-Q Plots (Quantile-Quantile plots) are two quantiles' plots 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. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. . 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.

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.

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.