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?

Categorical columns such as "instant, dteday,casual and registered" are not useful for analysis. Rest of the categorical columns such as “season,weathersit,mnth,weekday,workingday,holiday” are useful for analysis and their effect on target variable is as below-observed from Boxplots plotted against target variable “cnt”:

workingday: Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years) which is an indicator of how workingday can be a good predictor for the dependent variable.

mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

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

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.

Let’s assume we have three types of values under categorical column “Sex”- Male,Female,Others. Therefore, it’s understood that if one variable is male,2nd variable is female then obviously 3rd variable will be Others. So, it won’t be required to create a 3rd variable while creating dummies for this column.

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

We can observe from the pair-plot that “temp and atemp” variables have the highest correlation with target variable cnt.

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

The below assumptions of Linear Regression were validated after building the model on training set:

Error terms are normally distributed with mean zero (not X, Y)- During Residual analysis we plotted a Histogram and observed that Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

Linear Relationship b/w X&Y- Using the pair plot, we could see there is a linear relation between temp and atemp variable with the predictor 'cnt'

No Multicollinearity between the predictor variables- From the VIF calculation we could find that no multicollinearity exists between the predictor variables, as all the VIF values are within the permissible range of below 5.

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

As per the final Model 6, the top 3 predictor variables that influence the bike rental bookings are:

Temperature (temp) - A coefficient value of '0.5636' indicated that a unit increase in temp variable increases the bike hire stats by 0.5636 units.

Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire stats by 0.3070 units.

Year (yr) - A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire stats by 0.2308 units.

Therefore, it's suggested to give these predictor variables the utmost importance while booking bike rentals to achieve maximum booking.

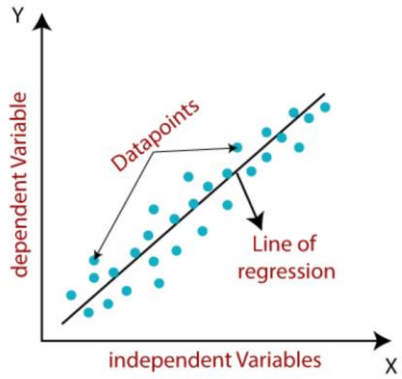
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Fit line equation $y= a_0+a_1x+ \varepsilon$

Here,

Y=Dependent variable

X=Independent variable

a0=intercept of the line

a1=linear regression co-efficient

e=random error

The values for x and y variables are training datasets for Linear Regression model representation

Types of Linear Regression

Simple Linear Regression- If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple linear Regression- If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

Positive Linear Relationship- If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

Negative Linear Relationship- If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line, so to calculate this we use cost function.

Cost function-

- The different values for weights or coefficient of lines (a_0, a_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

$$MSE= 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1x_i + a_0))^2$$

Where

N=Total number of observation
Yi = Actual value
$(a1x_i+a_0)$= Predicted value.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent:

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

1. R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
- It can be calculated from the below formula:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

Linear relationship between the features and target: Linear regression assumes the linear relationship between the dependent and independent variables.

Small or no multicollinearity between the features: Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.

Homoscedasticity Assumption: Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

Normal distribution of error terms: Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
It can be checked using the **q-q plot**. If the plot shows a straight line without any deviation, which means the error is normally distributed.

No autocorrelations: The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

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 analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze 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 behavior irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
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

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

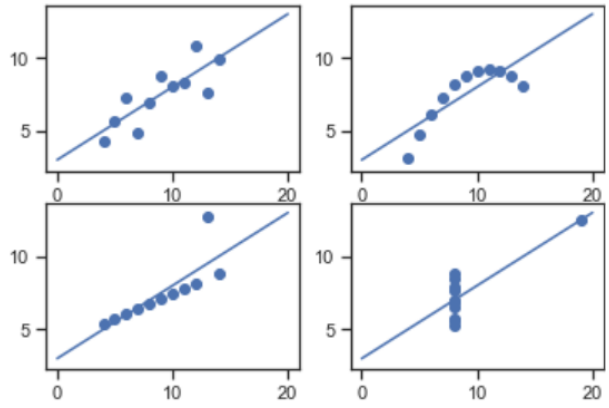
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



Graphical Representation of Anscombe's Quartet

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

3.What is Pearson's R?

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation

- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

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

Scaling is the process to normalize the data within a particular range. In linear regression, scaling is often recommended to make it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their means. Scaling is required when multiple variables are in different ranges. The two most discussed scaling methods are Normalization and Standardization. Scaling helps the model to find the real insights about the data.

In normalization, we map the minimum feature value to 0 and the maximum to 1. Hence, the feature values are mapped into the [0, 1] range: In standardization, we don't enforce the data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1: It not only helps with scaling but also centralizes the data.

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

A large value of VIF indicates that there is a correlation between the variables, and if there is perfect correlation, then VIF = infinity. An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model. This happens when R-squared value is equal to 1, which makes the denominator of the VIF formula 0 and the overall value infinite.

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 type of plot that we can use to determine whether or not the residuals of a model follow a normal distribution. If the points on the plot roughly form a straight diagonal line, then the normality assumption is met. Q-Q plots are used to determine if two data sets come from populations with a common distribution. In linear regression, Q-Q plots are used to check if the residuals are Gaussian and thus the errors are as well.