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)

Categorical variable	Effect on the dependent variable
season	The bike usage in descending order - fall, summer, winter and spring. This is kind of intuitive – drier & less harsher the weather more the usage.
weathersit	The bike usage in descending order - clear, cloudy and lightrain. This is also intuitive – usage is less if it rains
year	The bike usage is more in the 2nd year (2019) than in the 1st year (2018)
month	The bike usage is more in the middle of the year (which coincides with the fall season) than in the beginning or end of the year
weekday	The typical bike usage is same over all the days in a week. But most users seem to use the bikes on Thursdays and Fridays
workingday	The bike usage is not very different on working or non-working days
holiday	The typical bike usage is more on a non-holiday than on a holiday

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

It is possible to encode the (n) various values of a categorical variable using n-1 columns (with 0 or 1 values). If we use n columns, 1 column will be redundant and will introduce collinearity among the dummy variables.

For example, consider the categorical variable weathersit with values – clear (1), cloudy(2) and lightrain (3). You can encode these 3 values using 2 dummy variables. See below:

	cloudy	lightrain
clear	0	0
cloudy	1	0
lightrain	0	1

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

Registered users has the highest .95 correlation with the target variable

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

Assumptions of Linear Regression:

Assumptions	Validation
There is a linear relationship between X and Y	Using pairplots. Check the linear relationship between independent variables and the dependent variable.
Error terms are normally distributed with mean zero(not X, Y)	Calculated the residuals (ytrain - ytrain_pred) and plotted the distplot. Check that the residual plot is normal around the mean = 0.

Error terms are independent of each other	By plotting y_pred and residuals and check there is no pattern in the errors
Error terms have constant variance (homoscedasticity)	By plotting a scatter plot between the y_pred and residuals and check that the variance should be constant

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

Features	**Explanation of the model**
registered	More the number of registered users, more is the demand for the shared bikes
temp	Higher the temperature, more is the demand for the shared bikes
Mon	The demand for the shared bikes is higher on Mondays.

General Subjective Questions

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

Linear regression is a statistical method that allows us to summarize and study relationships between continuous (quantitative) variables.

- Simple Linear regression deals with 2 continuous variables – 1 predictor/independent variable and 1 response/dependent variable. The model fits a line
- Multiple Linear regression deals with multiple continuous variables – n (>1) predictor/independent variable and 1 response/dependent variable. The model fits a hyperplane instead of a line.

Note that the Linear Regression establishes an association relationship and not a causal relationship

The relationship between the variables is represented by the following equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... \beta_p X_p + \epsilon$$

where

- $\beta_0, \beta_1, \beta_2$… are estimated regression coefficients. β_0 is the intercept. β_1 is the slope
- ϵ = is the prediction error/residual error

The aim of Linear Regression is to find a **Best-fitting line** a line that fits the data "best" i.e. the line for which the prediction errors are as small as possible. We need to find the values $\beta_0, \beta_1, \beta_2$…that make the sum of the squared prediction errors the smallest or get the "least squares estimates" for $\beta_0, \beta_1, \beta_2$…

Following are the metrics for evaluating the model

- **Residual Sum of Squares**: Quantifies how much the predicted data points vary around the estimated regression line. It is the variation in Y that is not explained. Smaller the value of RSS, better is the model
- **R-squared** is the variation in Y that is explained by the model to the total variation in Y. So higher value of R-squared is better

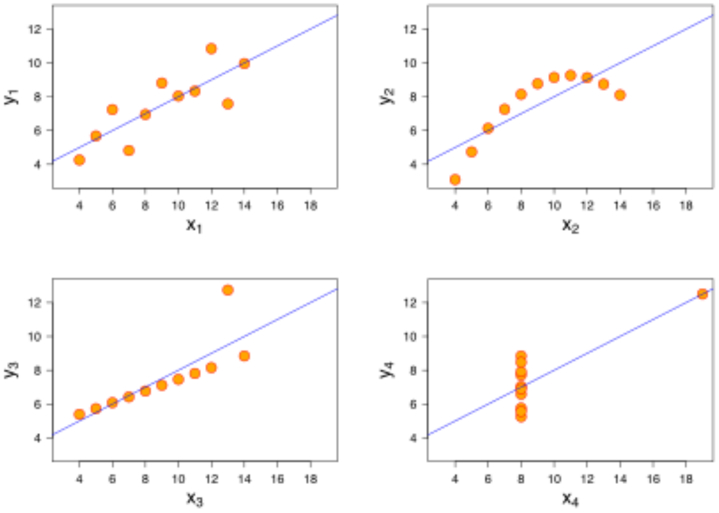
Assumptions of Linear Regression: A linear regression model should comply to the following assumptions.

- There is a linear relationship between X and Y
- Error terms are normally distributed with mean zero(not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

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

Anscombe's quartet comprises of four data sets that have nearly **identical simple descriptive statistics**, yet have **very different distributions** and appear very different when graphed. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate

- the importance of graphing data before analysing it
- the effect of outliers and other influential observations on statistical properties.



Data-set I	Consists of a set of (x,y) points that represent a linear relationship with some variance
Data-set II	Shows a non-linear relationship
Data-set III	Shows a tight linear relationship between x and y, except for one large outlier
Data-set IV	Value of x remains constant, except for one outlier as well

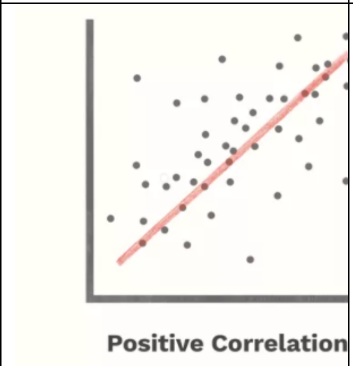
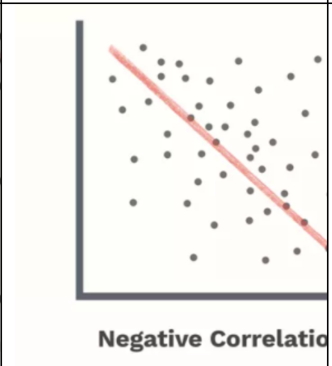
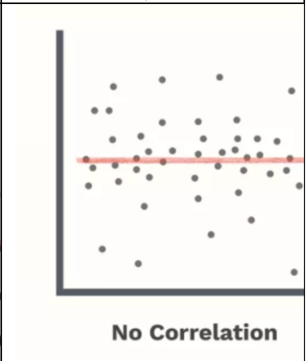
Anscombe's quartet was intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough". The quartet is often used to **illustrate the importance of looking at a set of data graphically before starting to analyse and the inadequacy of basic statistic properties for describing realistic dataset**.

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

Pearson's R is a correlation coefficient used in Linear Regression. Correlation coefficient is a measure of **how strong a (linear) relationship is between two continuous variables**. Note that it a measure of **linear** correlation of variables.

It takes a value between -1 and 1.

- The absolute value of the correlation coefficient gives us the relationship strength or magnitude. The larger the number, the stronger the relationship.
- The sign indicates the direction of the relationship i.e. with every positive increase in one variable if there is an increase/decrease in the value of the other variable

Value & Relationship	1 indicates a strong positive relationship	-1 indicates a strong negative relationship	A result of zero indicates no relationship at all
Plot			
How to interpret ?	For every positive increase in one variable, there is a positive increase of a fixed proportion in the other.	For every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.	There isn't a positive or negative increase.
Example	Shoe sizes go up in perfect correlation with foot length.	Amount of gas in a tank decreases in perfect correlation with speed.	There is no relation between the height of a person and the place where the person stays.

Python – functions that can be used to calculate and visualise correlation

- seaborn.scatterplot : Visual the linear relationship and the strength of the relationship.
- pandas.DataFrame.corr : Compute the pairwise correlation of variables in a dataset
- Seaborn.pairplot: Visualise pairwise correlation of variables in a dataset
- Seaborn.heatmap: Visualise correlation matrix as heatmap

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

The concept of scaling is relevant when continuous independent variables are measured at different scales. It means these variables do not give equal contribution to the analysis. The coefficients of the resultant model will also differ in scale and this would impact the interpretability of the model.

For example, we are performing customer segmentation analysis in which we are trying to group customers based on their homogenous (similar) attributes. A variable called 'transaction amount' that ranges between $100 and $10000 carries more weightage as compared to a variable i.e. number of transactions that in general ranges between 0 and 30. The coefficient for transaction amount might be smaller (like in the range 1 or lesser) whereas the coefficient of the number of transactions would be higher (10 or more). Based on the high value of the coefficient, we cannot conclude that number of transactions is more important Hence, it is required to transform the data to comparable scales. The idea is to rescale an original variable to have equal range and/or variance.
Scaling helps in the following ways:

- It can make the analysis of coefficients easier. If your features differ in scale then this may impact the resultant coefficients of the model and it can be hard to interpret the coefficients.
- It can help in the faster convergence of the algorithm in case you are using Gradient Descent.

Normalized or MinMax Scaling	Brings all the data in the range of 0-1	x = [x - min(x)]/[max(x)-min(x)]
Standardised scaling	Brings all the data into a standard normal distribution with mean 0 and standard deviation 1	x = [x-mean(x)]/sd(x)

The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there is are extreme data point (outlier)

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

If the VIF value comes out to be huge like infinity it means that the variable has perfect correlation with other independent variables.

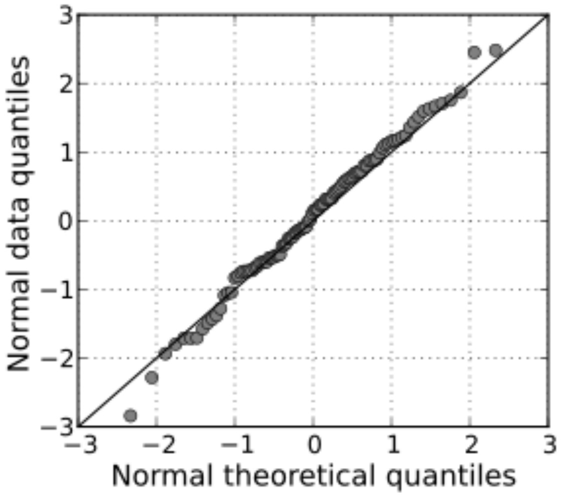
In such cases, the value of R2 will be 1 and VIF = 1/ (1-R2) = 1/ (1-1) = 1/0 = infinity

To solve this problem, drop one of the variables that is causing multicollinearity. While dropping variables, see that variables that do not help in interpretation of the model are dropped.

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

The Q-Q plot, or 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 or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. But note that this is just a visual check, not an air-tight proof.

Q-Q plots compares (using a scatter plot) the quantiles of our against quantiles calculated from a desired (theoretical) distribution. If both sets of quantiles came from the same distribution, the points form a line that’s roughly straight. An example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



While **Normal** Q-Q Plots are the ones used most often, these can be used for **any distribution** - exponential or uniform distribution

Q-Q plots can be used to verify if the **normal error term** (that requires the errors/residuals to be normally distributed) assumption of linear regression is met or not.