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)

- The below inferences can be drawn from the categorical variables analysis :
  - Fall season attracted the highest booking followed by summer and spring season attracted the least bookings.
  - The booking count is increasing year on year which signifies positive trend in consumption of bike taxis.
  - The number of booking is highest in the mid-year month's i.e- June till October, the trend is decreasing from November till Jan and then continues to increase.
  - Clear weather attracted more bookings than other types which is obvious.
  - During holiday's the average booking seems to be little than non-holiday's which signifies people could be possibly using the service to commute to work.
  - The number of booking doesn't get much affected by the day of the week.
  - The number of booking doesn't get much affected if day is a working day or not.

The overall conclusion can be drawn as we are observing positive trend in the service consumption and considering the above factors we can scale the business to next level.

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

- We use the drop_first=True during dummy variable creation to create n-1 unique values of the column from n values, it reduces the correlations while creating dummy variables. Suppose we have a column with n distinct values, the drop_first=True while dummy variable creation will n-1 which remove nth value which correlation with itself.
  Syntax: pd.get_dummies({Dataframe Name}.{categorical column name},drop_first=True)
  e.g- pd.get_dummies(df.season,drop_first=True) where df is the dataframe name, season is the categorical column name.

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

- The variable 'temp' has the highest correlation with the target variable as inferred from the pair-plot among the numerical variables.

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

- I have validated the assumptions of Linear Regression after building the model on the training set based on the below assumptions:
  - Normality of error terms
    The error terms are normally distributed in the model built.
  - Multicollinearity check

From the heatmap, I have observed that there is insignificants/least multicollinearity among variables.

- Homoscedasticity
  There is no visible pattern in residual values as observed from the scatter plot.
- Linearity validation
  We can observe linearity among finally selected features/variables by the model.
- Independence of residuals
  We can see the residuals has independence and not correlated to each other's.

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

- The below are the top 3 features contributing significantly towards explaining the demand of the shared bikes :
  - temp
  - winter
  - summer

General Subjective Questions

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

- Modelling uses machine learning algorithms, in which the machine learns from the data justlike humans learn from their experiences. Machine learning models can be classified into the following three types based on the task performed and the nature of the output:

  1. Regression: The output variable to be predicted is a continuous variable, e.g. scores of a student

  2. Classification: The output variable to be predicted is a categorical variable, e.g. incoming emails as spam or ham

  3. Clustering: No predefined notion of label allocated to groups/clusters formed, e.g. customer segmentation for generating discounts.

  Regression and classification fall under supervised learning methods – in which you have the previous years' data with labels and you use that to build the model.
  Clustering falls under unsupervised learning methods – in which there is no predefined notion of labels.

  Linear regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). There are two types of linear regression:
  ● Simple linear regression
  ● Multiple linear regression

  Mathematically the relationship between dependent (target variable) and independent variables (predictors) can be expressed by help of following equation

  $Y = B_1X + B_0$, where
  Y is the dependent variable
  X is the independent variable
  B_1 is the slope of the regression line
  B_0 is a constant/Y-intercept for X = 0, Y would be equal to B_0 .

  The below are assumptions are made by Linear Regression model:

  - Normality of error terms:
    There should be a normal distribution of error terms with mean equal to zero.
    The assumption of normality is made, as it has been observed that the error terms generally follow a normal distribution with mean equal to zero in most cases and can be validated by plotting a histogram.
  - Multicollinearity check
    The variables should be independent of each other and there should no correlation between the independent variables.
    This can be validated using VIF scores<5 or by checking the correlation matrix or heatmap of it.

- Homoscedasticity

  The variance of error terms should be constant i.e the spread of residuals should be constant for all values of X. This can be validated using scatter plot and there should be no visible pattern in residual values to be observed from the scatter plot.

- Linearity validation

  It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.

- Independence of residuals

  We can see the residuals have independence and not correlated to each other's.

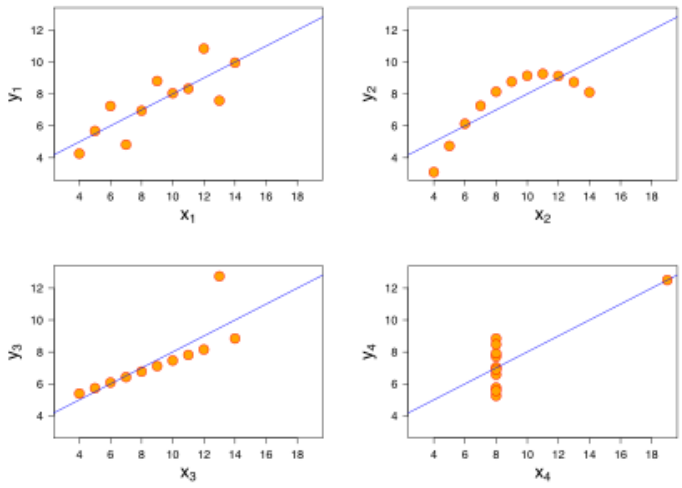
2. Explain the Anscombe's quartet in detail. (3 marks)

➢ The Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

The Anscombe's quartet emphasis on importance of visualization in Data Analytics and Data science, by looking at the data alone which looks like that should have similar characteristics is not correct, when we visualize the data, it will reveal the true nature and fit for the Machine Learning algorithms.

Anscombe's quartet								
STATS	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	82.51	99	82.51	99	82.5	99	82.51
AVG/MEAN	9	7.5	9	7.5	9	7.5	9	7.5
STD DEV	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Consider the above dataset, the various statistical values remains the same for all the four datasets(quarters), while the visualization reveals a different story and makes only 1 quadrant (quartet-I) fit for linear regression, which can be observed from the below scatter plots.



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

- 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.
  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.

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 0-1	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.

Reference: https://www.scribbr.com/statistics/pearson-correlation-coefficient/

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

- Feature Scaling is a technique to standardize the independent features(X variables) present in the data in the range. It is performed during the data preparation step to handle highly varying 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, irrespective of unit of the values.

  The key differences between normalized scaling and standardized scaling were as follows:

- In normalized scaling Minimum and maximum value of features are used for scaling while the other used Mean and standard deviation is used for scaling.
- In normalized scaling is used when features are of different scales while the standardized scaling is used when we want to ensure zero mean and unit standard deviation.
- In normalized scaling scale values between [0, 1] or [-1, 1] are used for scaling while the standardized scaling is not bounded to a range.
- In normalized scaling is affected by the presence of outliers while the standardized scaling is much less affected by the presence outliers.
- For normalized MinMaxScaler to be used while the standardized scaling will need StandardScaler.

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

- A VIF value of infinite signifies perfect correlation or extremely strong correlation between the variables. This mean multicollinearity exits among the selected features in the dataset.We get RSqaure value=1, then 1/(1-Rsquare) becomes infinity.
  We would need to find and drop the features which has very high correlation, this would help us to remove/reduce the multicollinearity.

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 plots are also known as Quantile-Quantile plots and plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

Use of q-q plot:

- The Q-Q plot permits identification of any peculiarities of the shape of the sample distribution, which might be symmetrical or skewed to higher or lower values.
- If the two datasets (samples) are taken from a population with the same distribution, the points should fall approximately along the reference line of population. The greater the deviation from the reference line, the greater the chance that the two data sets are from populations with different distributions.

Importance of Q-Q plot:

- 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.
- It can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic.