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)

To analyze the categorical column, I used the box plots and bar plots to assess the relationship between the categorical columns and the dependent column i.e. count

Below is my finding,

- **Season** –
  - The maximum number of bookings are made in **fall** season which indicates
  - The number of booking increases from spring < summer < fall and then reduces in winter.
  - The booking for each season is increasing year by year.
- **Month** –
  - The highest number of booking is done in month of **September.**
  - The number of bookings are increasing from Jan->Sept and then decreases month by month till December.
  - The number of bookings for each month is increasing year by year.
- **Weekday** –
  - The most number of bookings are made on Fri, Sat
  - This clearly indicates that the customer is opting to make the booking on weekends.
- **Holiday/Workday –**
  - Most number of bookings are made on Holidays in comparison to Workdays.
  - People tend to make lesser bookings on workdays
- **Year**
  - The number of booking is increasing year by year

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

When the dummy variables are created on the categorical columns, For Each Categorical Column where k levels are present, k dummy columns are created. These dummy columns are highly correlated with each other and to reduce this correlation, the first column is suggested to be dropped. This dropped column can be easily derived using the rest of the columns since these are correlated.

So if we use drop_first = false

For k levels, k dummy columns will be created

If we use drop_first= true,

k-1 dummy columns will be created as the first column will be dropped

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

Looking at the pair plot, The columns 'temp', 'atemp' are highly correlated with the target column count.

Since 'temp' and 'atemp' are also showing very high correlation i.e. 99%, we have dropped 'atemp' column.

Hence we can say that 'temp' column is highly correlated with 'count' target column.

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

The validation of assumptions are done using the 5 assumptions,

1. **The Normality of Errors**
    - This was achieved by plotting the residuals (y train – y pred) using a distribution plot where the mean was centered to the value of 0 and the distribution is following a normal distribution.
2. **The Multicollinearity**
    - This was achieved by plotting a heatmap on the training set data and there is no strong collinearity between two independent columns. Industry standard of 70% + is considered.
3. **Linear Relationship**
    - By giving an example of CCPR, The 'temp', 'wind-speed' columns are showing linear relationship.
4. **Homoscedasticity**
    - This is achieved by plotting a scatterplot between residual and count column (Target).
    - There is no visible pattern in residual values
5. **Independence of correlation**
    - There is no auto correlation

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

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

1. Temp = 0.432
2. Light_snowrain = -0.288
3. Year = 0.235

General Subjective Questions

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

Linear regression Algorithm is a statical model that is used on the continuous variables and which analyses the relationship between the target variable with the other independent variables.

The relationship between the independent variable vs target variable can be in both positive and negative.

This relationship can be mathematically represented with below equation,

Y=mx +c

Where m = coefficient of x

C = constant

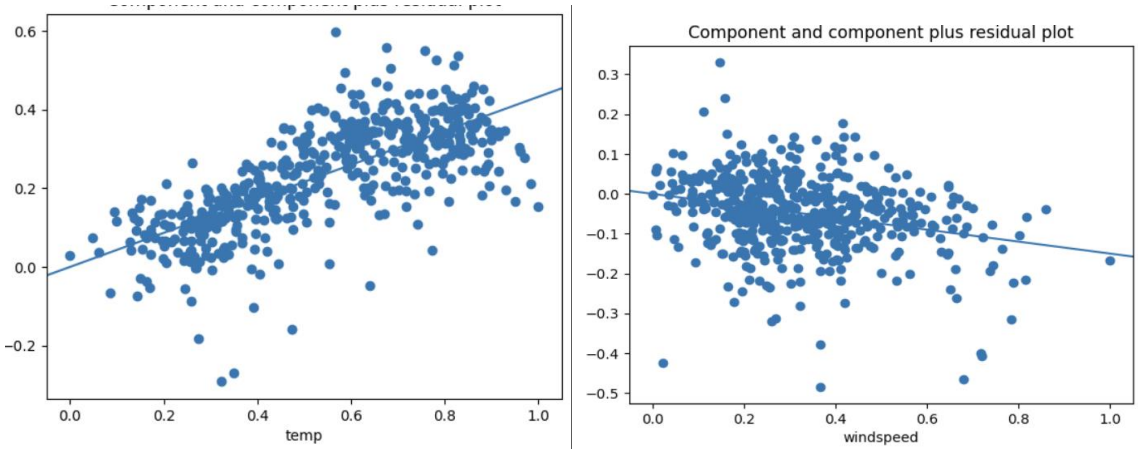
Y = target variable

X = independent variable

The Relationship can be both positive and negative.

Below example shows both the relations using the same shared bike assignment

There is a positive relationship between temp and count variables and also there is a negative relationship between the windspeed and count variable as shown below



There are two types of Linear regression:

1. Simple Linear Regression – **SLR**
2. Multiple Linear Regression – **MSR**

There are assumptions that are followed:

1. **Multicollinearity** – There should be no collinearity between the independent variables.
2. **Linear Relationship** – The independent variables must show linear relationship with target variables.
3. **Normality of Errors –** The residuals must be normally distributed and must be centered at mean at 0.
4. **Auto Correlation** – Auto correlation happens when there is a correlation between the residual errors
5. **Homoscedasticity** – There should be no visual pattern between the residuals values.

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

The Anscombe's quartet is a group of data set consisting same mean, standard deviation and regression line.

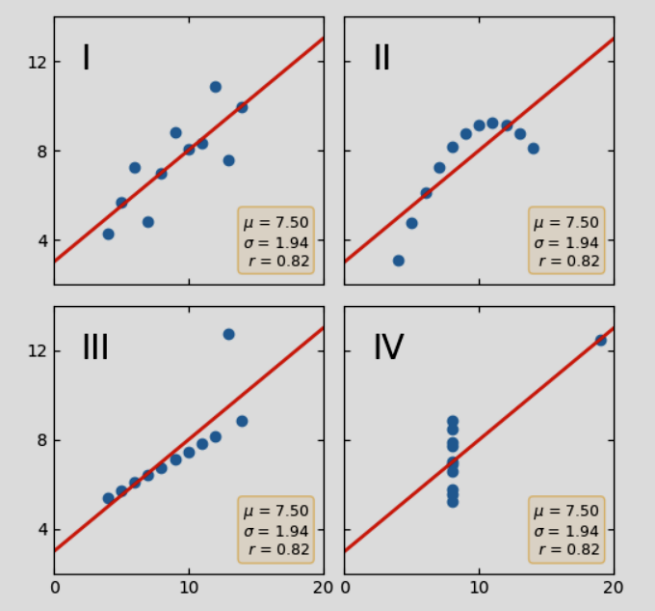
The data is used to illustrate that we should not rely on just the basic statical properties of the dataset, instead we should also look at the visual representation of the dataset.

The Anscombe's quartet shows 4 different dataset that results in same mean, std, regression line.

Below is the dataset,

Red		Blue		Yellow		Green	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Below is the visual representation of the same dataset.



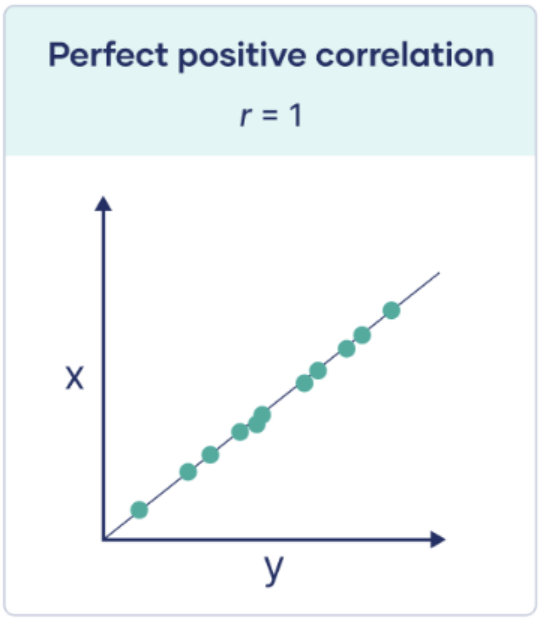
- Dataset 1 appears to have a well fitting linear models
- Dateset 2 is not distributed normally
- Dataset 3, the distribution is linear but the calculated regression is thrown off by an outlier
- Dataset 4, shows that even one outlier is enough to produce a high correlated coefficient.

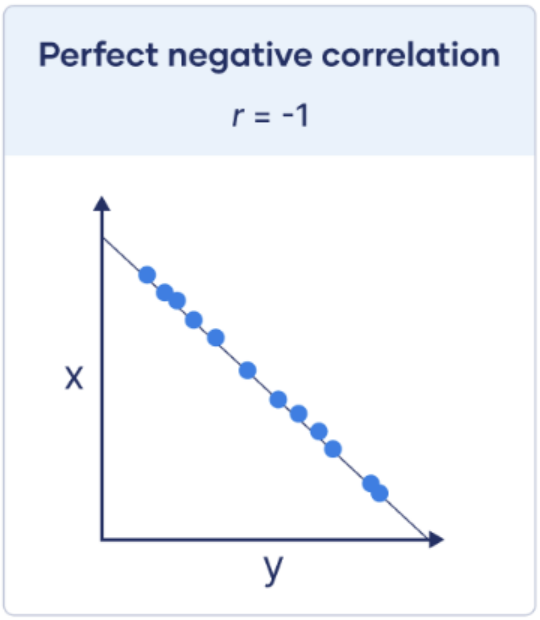
3. What is Pearson's R? (3 marks)

Pearson's R is a way to measure a linear correlation.

It is denoted by a number between -1 and 1 that measures the strength of correlation between two variables where as the positive or negative sign denotes the direction of the relation.

- If the number is between 0 to 1, It denotes a positive correlation. The change in one variable is directly proportional to other variable in positive direction.
- If the number is between 0 and -1, then it denotes a negative correlation and the change in one variable is inversely proportional to other variable.





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

Scaling is a technique to standardize the independent variables in a fixed range.

It is performed to unify the unit and standard of all the variables. This is performed in the pre-processing step so that the correlation between the independent variables and target variable results in correct coefficient.

S.NO.	Normalized scaling	Standardized scaling
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.

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

Whrn the correlation between two independent variable is perfect. A large value of correlation explains that there is a correlation between the independent variables.

The R squared becomes 1

Which leads to 1/(1-R2) as infinity

To solve this problem, we have to drop one of the correlated independent variable and use the other one in our analysis.

Dropping the correlated indenependent variable also solves the problem of multi-collinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   This is a plot of quantiles of the first dataset against the quantile of the second dataset.
   Quantile means the percentage of points below the given value.
   Example:
   0.4 or 40% quantile means the point at which the 40% of data falls below the value and 60% of the data falls above the value.

   If there are two data samples, Then its used to find out if the assumption of the common distribution is justified.

If the tow samples differ, it also helps to understand the differences.

It provides more insight on the nature of the differences than the analytical methods.