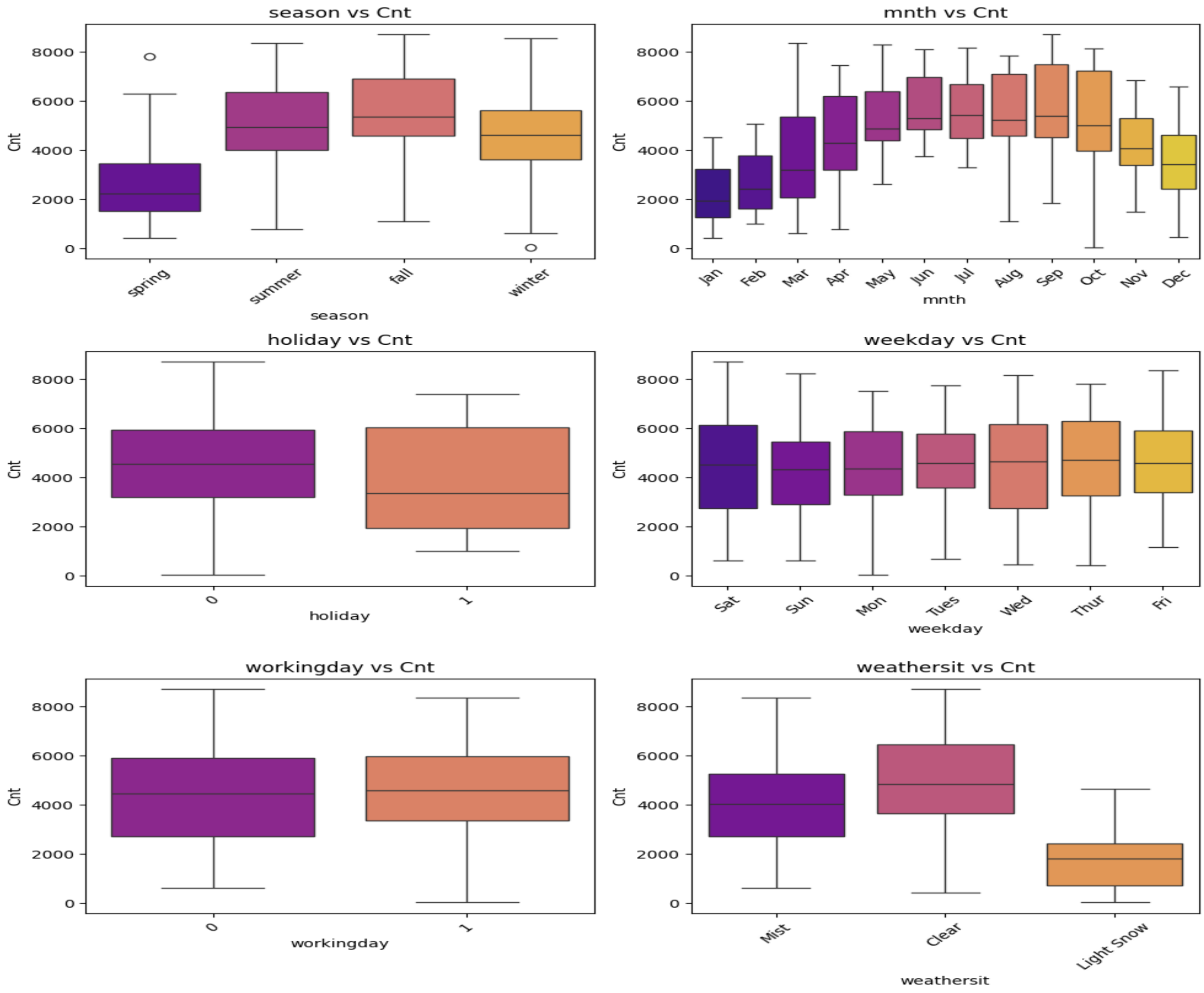
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Below are the inferences for categorical variables.

   a) On a holiday, demand decreases.
   b) Fall has the highest and spring seasons has the lowest bike rental bookings.
   c) Demand increases each month from Jan to June. The highest demand is in the month of September. Demand starts decreasing after September month.
   d) Demand is more when whether is Clear. It can be seen from the last chart.

The above findings can be verified from the chart below.



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

Dummy variables are created when we have categorical variables in our dataset. These dummy variables convert the categorical values into 0 and 1 values such that if there are n types of values in a single categorical column it will create n-1 variables. Drop_first=True instructs runtime to delete the nth variable when creating dummy variable such that final output is n-1.
A very common example is to explain the flat type column. Suppose you have a housing dataset where you have flattype column with 3 values furnished, semifurnished and unfurnished. In this case when we create dummy variables it will create n-1 i.e. 3-1 = 2 variables with below values which would be like below

Furnished	semifurnished
1	0
0	1
0	0

When both columns have zero value, it means flat is unfurnished.

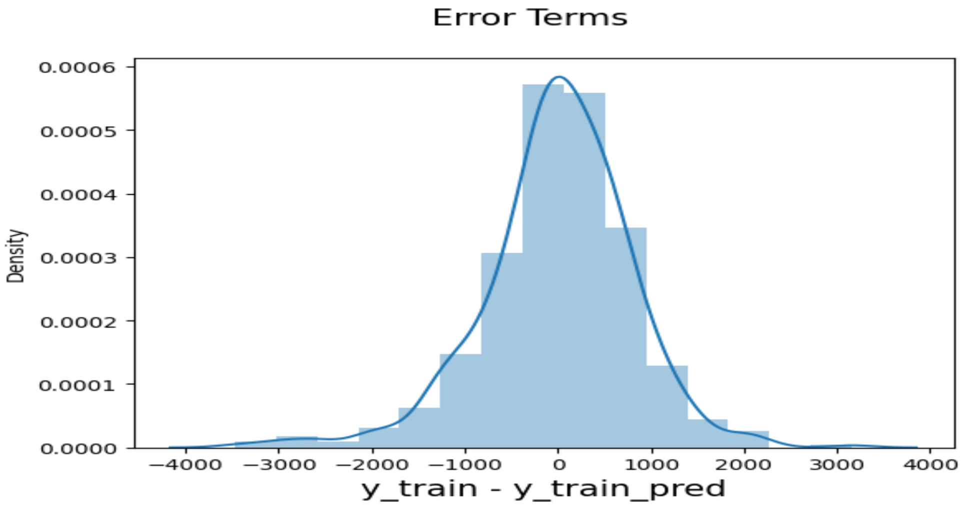
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Variables having highest correlation with cnt variable are Temp and atemp

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

a) By using below chart. We can see that the error terms are centered around zero. This proves residual distribution is a normal distribution.
b) Using VIF (Variance Inflation Factor) and p-values we verified assumptions of Linear Regression of multicollinearity. P-value greater than 0.05 indicates that the test result is non-significant. Verifying that there is linear relationship between independent and dependent variables.
c) By using a pairplot we observed which numeric variables have linear relationships with target variable cnt and which variables are highly correlated. Like temp and atemp variables are highly correlated.

Chart below for point a).



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 top 3 features contributing significantly to explaining the demand of the shared bikes are

   1. season,
   2. weather,
   3. temperature or year as well.

   a) From the charts we can see that the Fall season has the highest demand for rental bikes and spring season has the least demand for rental bikes.

   b) From the charts we can see that the demand for rental bikes increases when the weather is clear and demand decreases when the weather is bad (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds).

   c) As temperatures increase so does the demand for bike increases.

   d) We can also see from the charts that the year 2019 saw more demand for rental bikes which may be because year on year the demand is increasing.

General Subjective Questions

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

   Linear regression algorithm explains a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

   Independent variable is the cause. Its value is independent of other variables. Dependent variable is the effect. Its value depends on changes in the independent variable. When there is only one independent feature, it is known as Simple Linear Regression, and when there is more than one feature, it is known as Multiple Linear Regression.

   The equation for simple linear regression is:

$y = \beta0 + \beta1X$
where:
y is the dependent variable
X is the independent variable
$\beta0$ is the intercept.
$\beta1$ is the slope.

The equation for multiple linear regression is:
$y = \beta0 + \beta1X1 + \beta2X2 + \dots\ ..\ ..\ \beta nXn$

where:
Y is the dependent variable
$X1, X2, \dots, Xn$ are the independent variables.
$\beta0$ is the intercept.
$\beta1, \beta2, \dots, \beta n$ are the slopes
Linear regression algorithm is a type of supervised machine learning algorithm because the algorithm learns through the training data. Training data is past data with labels is used for building the model.

The slope indicates the rate of change in y per unit change in x. The y-intercept indicates the y-value when the x-value is 0.

E.g where we use Linear regression is as follows.
a) predicting house prices based on features like size, location, and number of bedrooms.
b) predicting the sales of the products.

Steps involved in Linear regression analysis are as follows:
a) Reading, understanding and visualizing data.
b) Preparing the data for the model (train-test split, rescaling) etc.
c) Training the model (OLS – Ordinary least square).
d) Residual analysis
e) Prediction and evaluation on the test set (R-squared, Adj. R-squared).

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

Because of Anscombe's quartet we came to know about the importance of visualization. Looking at data in terms of charts gives us more insights which we can understand sometimes just by looking at numbers and a few mathematical statistics formulas.

Anscombe's quartet is a set of four dataset where data from each dataset has same statistical properties as mean, variance, R-squared, correlation and linear regression line but when we plot the graphs for each dataset, we get different representations. So, it explains the importance of visualization while performing EDA.

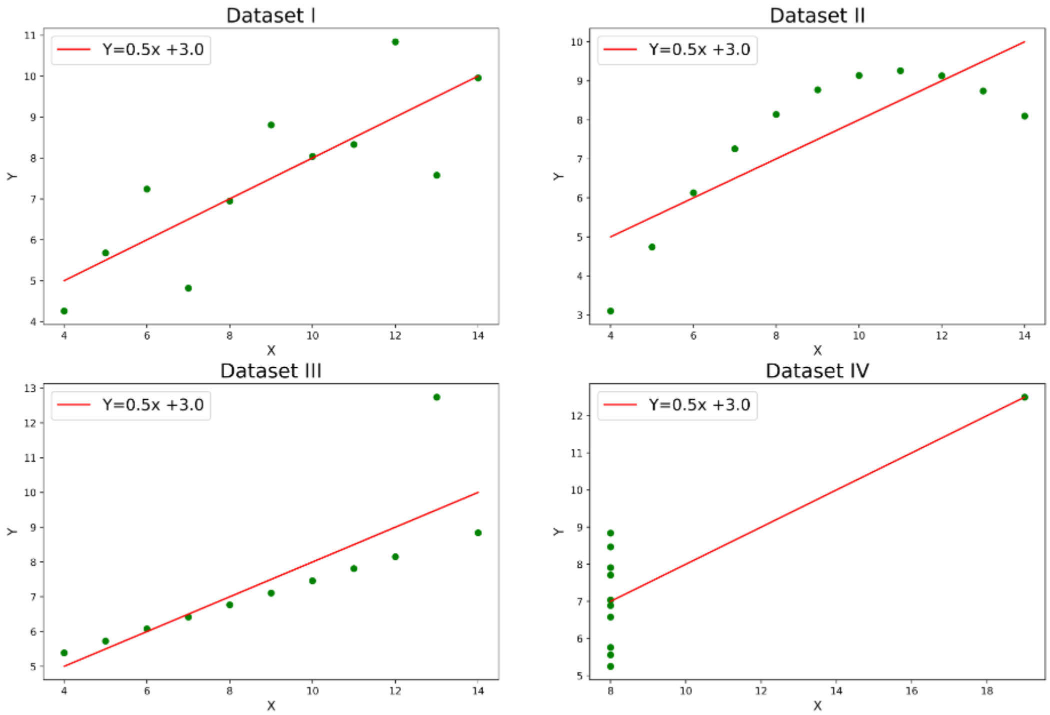
Below are the values for all four datasets. Source is Wikipedia.

Anscombe's quartet							
Dataset I		**Dataset II**		**Dataset III**		**Dataset 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

Below are different properties of the four datasets.

Property	**Value**	**Accuracy**
Mean of x	9	exact
Sample variance of x: s2	11	exact
x		
Mean of *y*	7.5	to 2 decimal places
Sample variance of *y*: s^2	4.125	±0.003
y		
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R2	*0.67*	to 2 decimal place

Below is the visualization of four datasets.



Dataset I show a clear linear trend.

Dataset II appears mostly linear but has an outlier that misrepresents the relationship.

Dataset III reveals a non-linear pattern that a simple linear regression would fail to capture.

Dataset IV shows how a single outlier can dominate the analysis and mislead interpretations.

Learnings from Anscombe's Quartet:

Importance of Data Visualization: Don't just rely on statistics values. Outliers values can significantly affect statistical analysis.

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

The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y.

Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship.

The Pearson coefficient is a measure of the strength of the association between two continuous variables.

e.g number of minutes or hours of running are positively correlated to number of calories burned.

In python we can calculate it using corr method.

```
df = pd.DataFrame(data)
r = df['RunningHours].corr(df['CaloriesBurned'])
print(r)
```

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

Scaling is the process of transforming data so that it fits within a specific range or distribution. In machine learning and statistical analysis, scaling is crucial because it can significantly affect the performance of algorithms.

When all the inputs to an algorithm are on the same scale, calculation is easy, and result can be understood easily. Scaling ensures that no variable dominates other variables in calculation.

E.g. In the housing example. We use MinMaxScaler to scale variables like 'area', 'bedrooms', 'bathrooms', 'stories', 'parking’, ‘price'. It is because bedrooms and bathrooms will have small integer values like 2,3,4 and so on. But the area will have values like 400,500,1200,3000 and so on.
So, to bring every variable on same scale we use scaling.

Normalized Scaling transforms the data to fit within a specific range, typically [0, 1]. It is used when features have different scales, and you want to compress them into a uniform range. It's especially useful for algorithms that do not assume a normal distribution.

Formula for Normalized Scaling is as below.

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)}$$

where x is the original value, min(X) is the minimum value in the feature, and max(X) is the maximum value.

Standardized Scaling transforms the data to have a mean of 0 and a standard deviation of 1. It is used for algorithms that assume data is normally distributed.

Formula for Standardized scaling is as below.

$$x' = \frac{x - \mu}{\sigma}$$

where x is the original value, μ is the mean of the feature, and σ is the standard deviation.

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

VIF is infinite when there is a perfect correlation between two independent variables. An infinite VIF suggests that the coefficients of the correlated variables are unreliable and can lead to misleading conclusions. Hence we remove variable with a VIF score of greater than 5.

VIF is calculated using below formula.

$$\text{VIF(Xi)}= \frac{1}{1-R^2}$$

If there is perfect correlation between two variables then R2 is near about 1 and in this case the denominator in above equation become 1 – 1 = 0. And 1 / 0 is infinity.

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

(3 marks)

Q-Q plot is a Quantile- Quantile plot. It is basically a graph where we can see our data and understand it in terms of distribution. Whether it is normal distribution or not. We check whether residuals are normally distributed in linear regression. Q-Q plot helps us to verify that using a chart.

It helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

After we are done with calculations of residuals we generally generate the Q-Q plot. Based on the analysis on Q-Q plot we either make changes in the data or perform further analysis.