Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

We have the following categorical variables in the given data set. ['season', 'year', 'month', 'holiday','weekday','workingday', 'weather_situation']
I have created box plots to understand the relationship between the categorical variables and the count and able to find the following inferences.

- "Season": The bike counts are observed in the Fall season, followed by summer season as favorable seasons for the business. In contrast, the lowest bike counts are observed in Spring season which is not favorable to bike business.
- "year": Bike rental counts are noticeably higher in year 2019 compared to year 2018. It implies the bike business is picked up more in 2019.
- "month": Bike counts increase steadily from **January** to a peak in **September**, then decline in **November** and are lowest in **December**. It implies that people avoid bikes in winter season and prefer to ride in fall season.
- "Holiday" : Bike rental counts are lower on holidays compared to non-holidays. People want to stay at home in the holidays so don't want to use bikes heavily.
- "Weekday" : Bike counts are relatively consistent across weekdays but slightly lower on **Saturday** and **Sunday** (weekends). People want to stay at home at weekends so don't want to use bikes heavily.
- "Working day" : Bike counts are slightly higher on working days compared to non-working days . It implies that people are using bikes to commute to their workplace.
- Weather Situation : Bike counts are higher in clear weather and lowest in "Light Snow rain" conditions. It implies that people are using bikes only in friendly weather conditions and avoids them in adverse weather conditions.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

When you have a categorical variable with n levels, you will get created n number of variables if you don't use drop_first = True. However, we require only n-1 number of variables to explain n levels of the categorical variable. If we don't drop one column here then that column will create multi-collinearity issue as other n-1 variables can explain the same relation.
For example: Let's say you have a categorical variable called "Color" with three categories: Red, Blue, and Green. When you create dummy variables, you get three binary columns:
Color_Red
Color_Blue
Color_Green
If you use drop_first=True, one of these columns will be dropped to avoid multicollinearity. Let's drop "Color_Red". The resulting dummy variables will be:
Color_Blue
Color_Green
When these two variable Color_Blue and Color_Green are false then it automatically implies that the variable is stating Color_Red.
So we dont require Color_Red variable in this case to avoid mult-collinearity.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Looking at the pair plot among all numerical variables, temp and atemp are having highest correlation against targe variable – count. **Their correlation value is 0.65 against target variable**

And these two variables are having highest correlation of 0.99 which implies that one of them should be eliminated to avoid multi-collinearity.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

The following assumptions are validated after building linear regression model.

1. **Linearity** :
   Checking the linear relationship between predictors and target variables.
   My model R Square and Adj R- Square values are around 0.83 which confirms that linearity exists between the predictors and target variables.
   And also, when I created scatter plot between fitted values and residuals and I could see no clear pattern which Indicates linearity.
2. **Independence of residuals:**
   Residual should be independent and should not have any relationship.
   I created a plot to understand the residuals behavior over time series. And it does not have any pattern which implies that residuals are independent.
3. **Homoscedasticity :**
   When we check the relation between residuals and predictors then we should not have any pattern which indicates Homoscedasticity
   I created a scatterplot between fitted values and residuals to understand the shape. There is no funnel shape or any clear shape in the data. It implies homoscedasticity of the residuals

4. **Normality of residuals**
   Residuals should be normally distributed having mean at zero to indicate the normality of the residuals.
   I created a histogram for the residuals to understand the mean location and the shape of the distribution. My histogram confirms that residuals are normally distributed in shape and, in mean at Zero.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The following are the top 3 features contributing significantly towards the demand in shared bike business.

- Temp : When the temperature is favorable then bike demand will be increased
- Year : Bike demand is increasing every year based on the given data
- Weather situation: If weather situation is bad like adverse situations – Light Snow rains then the demand will be decreased.

General Subjective Questions

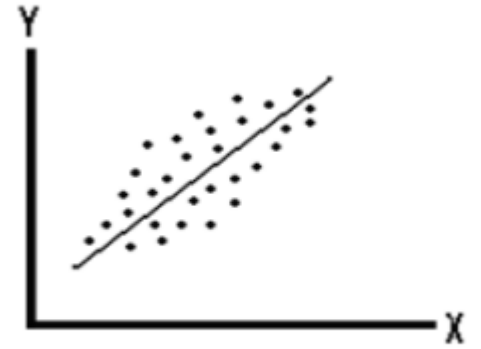
Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
Mathematically the relationship can be represented with the help of following equation —
$Y = mX+c$
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.
Furthermore, the linear relationship can be positive or negative in nature as explained below—

-

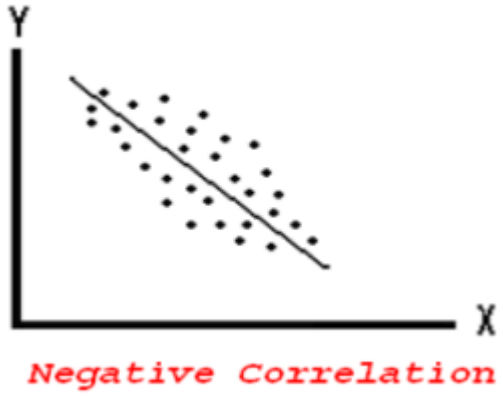
Positive Linear Relationship:
A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph



Positive Correlation

Negative Linear relationship:
A linear relationship will be called positive if

independent increases and dependent variable decreases. It can be understood with the help of following graph



Assumptions:
The following are some assumptions about dataset that is made by Linear Regression model
1. Multi-collinearity

Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

2. Auto-correlation
Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

3. Relationship between variables
Linear regression model assumes that the relationship between response and feature variables must be linear.

4. Normality of error terms
Error terms should be normally distributed

5. Homoscedasticity
There should be no visible pattern in residual values.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show the effect of outliers and other influential observations on statistical properties.

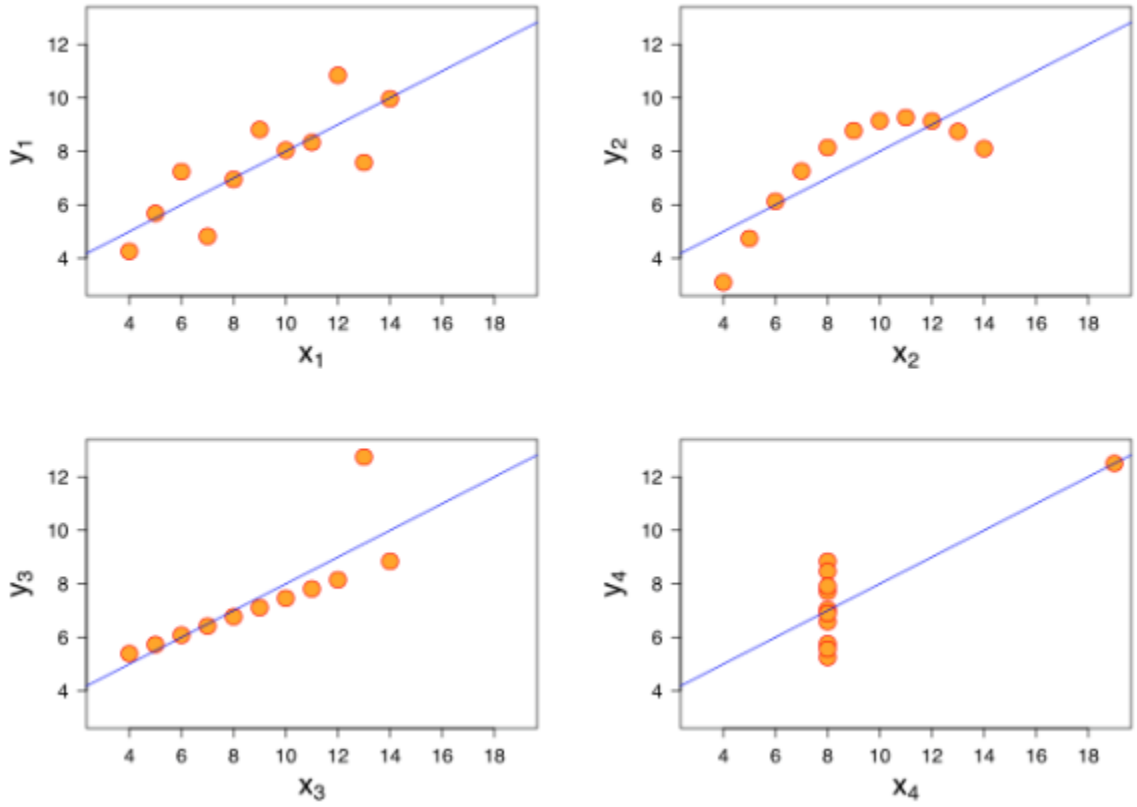
Anscombe's quartet							
Dataset I		Dataset II		Dataset III		Dataset IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	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

Key Points of Anscombe's Quartet

1. **Identical Statistical Properties**:
   - Each dataset has the same mean of x and y.
   - Each dataset has the same variance for x and y..
   - Each dataset has the same correlation coefficient between x and y.
   - Each dataset has the same linear regression line: $y = 3 + 0.5x$.
   - Each dataset has the same coefficient of determination R^2.

Visual Representation

Here are the four datasets plotted:



2. **Different Graphical Representations**:
   - **Dataset I**: Appears to be a simple linear relationship, fitting the linear regression model well.
   - **Dataset II**: Shows a clear non-linear relationship, indicating that a linear model is not appropriate.
   - **Dataset III**: Contains an outlier that significantly affects the regression line, demonstrating the impact of outliers.
   - **Dataset IV**: Has a high-leverage point that influences the correlation and regression line, despite the rest of the data showing no clear relationship.

Importance of Anscombe's Quartet

Anscombe's quartet highlights several important lessons in data analysis:

1. **Graphical Analysis**: Always visualize your data before performing statistical analysis. Graphs can reveal patterns, relationships, and anomalies that summary statistics might miss
2. **Effect of Outliers**: Outliers can significantly influence statistical measures and models. Identifying and understanding outliers is crucial for accurate analysis.
3. **Model Appropriateness**: The same statistical measures can suggest different models depending on the data distribution. It's important to choose the right model based on the data's graphical representation.
4. **Data Integrity**: Simple descriptive statistics can be misleading if not complemented with graphical analysis. This ensures a more comprehensive understanding of the data.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

The Pearson R also known as the 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.
The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between —l and 1 that measures the strength and direction of the relationship between two variables.
If the r value is between 0 to 1 its positively correlated and if r values is in between -1 and 0 then the variables are negatively correlated and If r value is 0 then there is no correlation.

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation.
The Pearson correlation coefficient is a good choice when all of the following are true:

1. **Both variables are quantitative:** You will need to use a different method if either of the variables is qualitative.
2. **The variables are normally distributed**: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
3. **The data have no outliers**: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
4. **The relationship is linear:** "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or 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, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 1000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Why Scaling required:

Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results into the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion is high. Also, the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

Standardization Scale:
Standardization (or Z-score normalization) transforms the data to have a mean of 0 and a standard deviation of 1. It centers data around 0 with a standard deviation of 1.
It is Used when the data follows a Gaussian distribution (normal distribution) and you want to compare scores across different scales.
Formula:

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

μ = Mean
σ = Standard Deviati

Normalization (or Min-Max scaling)
It rescales the data to a fixed range, typically between 0 and 1.
It is Used when you want to scale the data to a specific range, often between 0 and 1, especially when the data does not follow a normal distribution.

Formula:
m = (x — xmin)/(xmax — xmin)

m -> The new value
x -> The original cell value
xmin -> The minimum value of the column
xmax -> The maximum value of the column

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
<Your answer for Question 10 goes here>

VIF formula is $1/(1-R2)$
If there is perfect correlation, then VIF will be infinity.
When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared ($R2$) =1, which lead to $1/(1-R2)$

infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

Q-Q Plot checks the normality of the residuals in a given linear regression model.

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the quantiles of a dataset to the quantiles of a theoretical distribution, often the normal distribution. This helps in assessing whether the data follows a particular distribution.

#Use in Linear Regression

In linear regression, a Q-Q plot is typically used to check the normality of residuals. Residuals are the differences between observed and predicted values. For the assumptions of linear regression to hold, these residuals should be normally distributed. Here's why this is important:

Model Validity: Many statistical tests and confidence intervals in linear regression rely on the assumption that residuals are normally distributed. If this assumption is violated, the results of these tests may not be valid.

Detection of Outliers: Q-Q plots can help identify outliers or deviations from normality. Points that deviate significantly from the line suggest that the data may have heavy tails or other anomalies.

Improving Model Fit: If the residuals are not normally distributed, it might indicate that the model is not capturing all the patterns in the data. This can prompt further investigation and model refinement.

#Interpreting a Q-Q Plot

Straight Line: If the points in the Q-Q plot fall approximately along a straight line, the residuals are normally distributed.

S-shaped Curve: This indicates heavy tails, meaning there are more extreme values than expected in a normal distribution.

Inverted S-shaped Curve: This suggests light tails, meaning there are fewer extreme values than expected.
