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)

Important variables analyzed are shown below

1) Year – 2019 has more demand compared to 2018
2) Weather situation: Clear weather situation has higher demand, followed by mist and lowest demand for light rain
3) Working day and holiday: Working day has higher average demand compared to holiday
4) Season: Fall has highest demand and spring has lowest demand

---

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)

Drop_first = True will removes redundancy by eliminating unnecessary column while creating dummy variables. This helps in avoiding multicollinearity in linear regression model.

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)

Out of all numerical variables temperature(temp/atemp) has highest correlation with target variable. It has a correlation value of 0.63

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)

After building the model based on training data, following facts were checked

1. R2 score and adjusted R2 score – it is around ~ 0.85 , 85% of the cases can be covered
2. Test data set – 20-30% of data from original data set was passed to the model and R score checked, it is around 0.81. R2 score is close to training data set.
3. Residual analysis of test data and plotting of residuals – plot shows that the distribution stays closer to 0 indicating a good model.

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)

1. Temperature
2. Year
3. Weather situation

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>

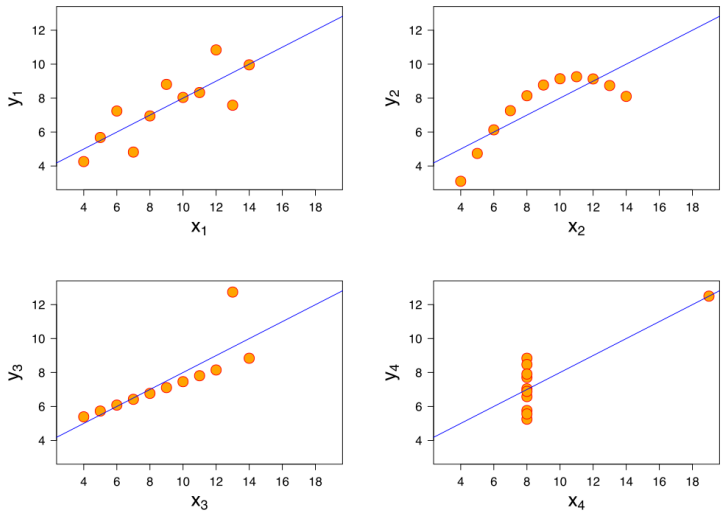
Linear regression is type of supervised learning used for predicting continuous dependent variable based on independent variables. There are two types of linear regression models. 1st one is simple linear regression – dependent variable has a correlation to only one variable. Multiple linear regression: dependent variable has correlation to multiple independent variables.
Formula of linear regression $Y=\beta 0+\beta 1X1+\beta 2X2+\ldots\ldots\beta nXn$

X1,X2..Xn are independent variables. Y is dependent variable. B0 is intercept when all X values are zeros.
Linear regression assumes that there is linear relationship between dependent and independent variables. Features should not be corelated. Errors are normally distributed.
Optimal line is derived by keeping the residual error to minimum.

Model is evaluated using R2 (coefficient of determination) – a statistical term that indicates how well the independent variables explain the variation of dependent variable. It ranges from 0 to 1. 1 is the perfect model.

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 comprises four datasets 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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough



Key Lessons from Anscombe's Quartet

1. Summary statistics are NOT enough

Even if mean, variance, and correlation are the same, datasets can be very different.

2. Always visualize your data

Scatter plots help in understanding actual relationships.

3. Linear regression is not always reliable

It fails for non-linear data and is sensitive to outliers.

4. Outliers can mislead analysis

One or two extreme values can distort correlation and regression results.

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>

In statistics, the Pearson correlation coefficient (PCC)[a] is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, 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. As a simple example, one would expect the age and height of a sample of children from a school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

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 the process used to transform numerical data into specific range (0-1 or -Sigma to +Sigma) making them suitable for machine learning algorithms

Scaling helps in four ways

1. Improves model performance
2. Optimizes the model
3. Reduces dominance of large variables
4. Improves stability

Normalized scaling fits the variables between 0 and one , it is also called as Min-Max scaling.
Standardized scaling has mean as 0 and it is normally distributed between standard deviation

---

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 is calculated by the formula 1/1-R2 , when R2 becomes zero VIF becomes infinity.
This is the case of high correlation. We need to drop some features with infinite VIF.

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>
The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.
They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Checking the Normality of Residuals
Linear regression assumes that residuals (errors) are normally distributed.
A Q-Q plot compares the quantiles of residuals with the quantiles of a normal distribution.
If points fall along the 45-degree diagonal, residuals are normally distributed.
If points deviate significantly, it suggests non-normal residuals, violating regression assumptions.
IT also helps in identifying outliers.
