Assignment-based Subjective Questions

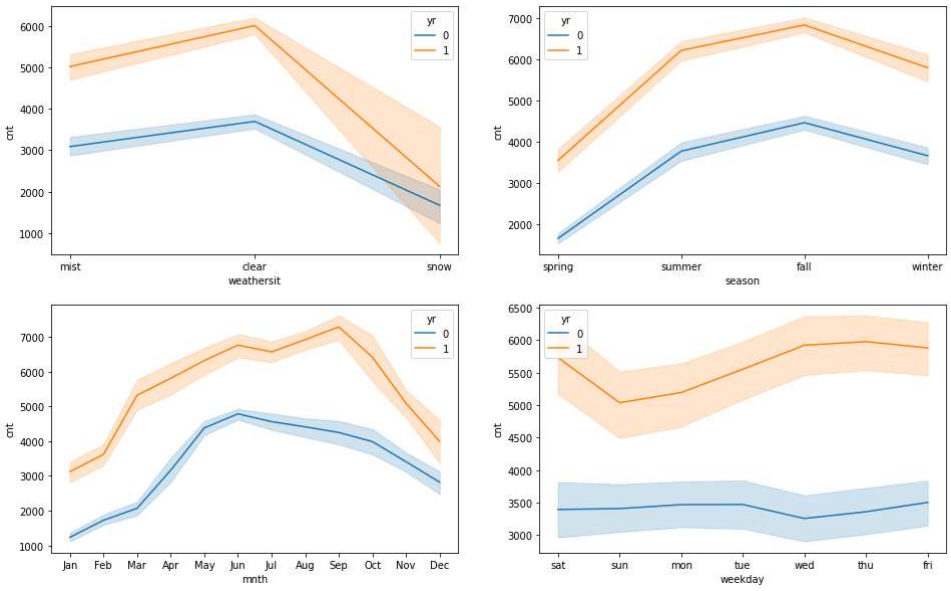
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

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

Clearly the categorical variables show influence on the demand for shared bikes.

- Demand is highest when the weather is good, followed by when misty and least when it is bad and snowy
- This also reflects in the demand by season where it peaks around summer and fall season while drops during winter and early spring
- There is no specific demand pattern during weekday or weekend.
- 2019 shows significant higher number of rentals than the previous year 2018 on all observations



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)

Using drop_first = True will effectively reduce an extra column per variable which would be redundant. It will further aid reducing multicollinearity among the created dummy variables and also improve the interpretability.

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)

The variables representing the temperature(temp) and feels-like temperature (atemp) have the highest correlation amongst other variables in the dataset.

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 dependent variable (cnt) holds a linear relationship with the independent variables e.g temp.
The residuals i.e. difference between predicted and observed value are normally distributed.
Minimal auto-correlation
The residuals have zero mean and constant variance
Homoscedasticity

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 temperature, year (we have only 2 years data) and the season appear to contribute most significantly to the demand of bikes.

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)

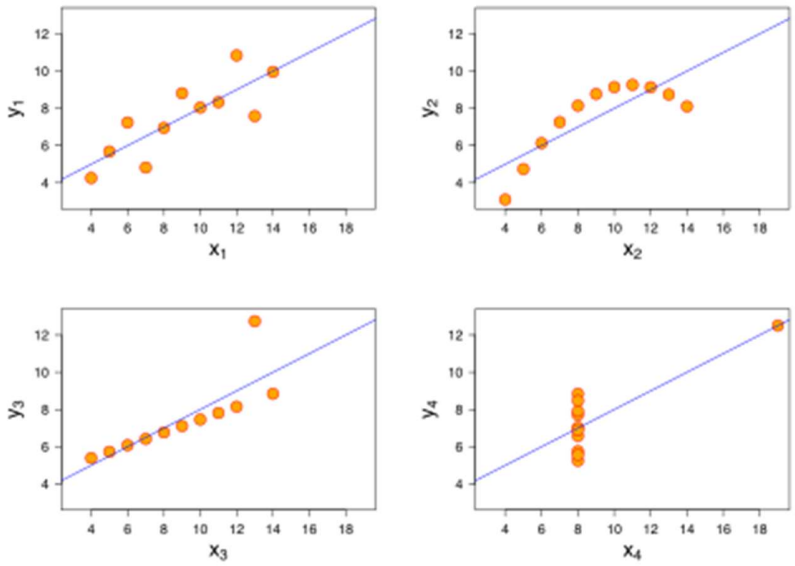
Linear regression is a type of supervised machine learning algorithm that learns from labeled data and identifies a linear equation, which can be used to predict on a new data. It effectively models the relationship between a dependent or target variable and one or more independent predictor variables (continuous).

Mathematically, the algorithm tries to find the best possible straight 'line' fitting the data by minimizing the sum of squared errors between predicted and actual values also called as the Least Squares method. The simplest one can be modelled as $y = mx + C$.

It assumes the data exhibits linear relationship between the dependent and all independent variables where the is minimal correlation between the variables. The models are considered valid and useful provided the errors show normal distribution and Homoscedasticity.

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)

Anscombe's quartet were constructed a named statistician Francis Anscombe to demonstrate the nuances on over-dependence on statistical and numerical methods, and the importance of using Visual data for analysis.
Anscombe constructed 4 datasets each with identical statistical parameters (mean, standard deviation and the correlation), but when plotted give a considerably different picture.



Dataset 1 shows linear relationship between x and y
Dataset 2 shows nonlinear relationship between x and y
Dataset 3 shows linear relationship with a outlier
Dataset 4 shows no relationship between x and y

All illustrate the importance of visually inspecting data when analyzing it, and the effect of outliers on statistical properties which could mis-lead the conclusions solely based on statistics.

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)

Pearson correlation coefficient (r) is a number between –1 and 1 to measure linear correlation. It indicates the strength and direction of the relationship between two variables. A high score (both positive and negative) indicates high similarity, while a zero indicates no correlation.

Mathematically the correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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)

Scaling is an important step while pre-processing or preparing data for machine learning model development specially when the data range and for independent variable varies widely thereby introducing sub-optimal results due to bias. For an example, few column values are in the range 1-10 while others are in 100-1000.

While both techniques are used to adjust the scale of variable values, below are the key

differences:

Normalization:

- Scales data to a set range e.g. 0-1 using techniques like **Min-Max** scaling
- Shape of original data in maintained in the scaled one
- It can be affected by outliers
- Used when the distribution is unknown

Standardization:

- Scales data to have a mean of 0 and standard deviation of 1
- Not bounded to a range
- Unlikely outlier impact
- Also called as Z-score and used when the distribution is Normal

---

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)

An infinite VIF or Variance Inflation Factor is an indication of perfect correlation between two apparently independent variables. The best way to address this is to drop one of the variables. The industry best practice is to keep the VIF below 5.

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)

The quantile-quantile (q-q plot) plot is a visual method for determining if a dataset follows a certain probability distribution or compare it with a known distribution like normal distribution.
It plots quantiles of the two data set where quantiles are points in a dataset dividing the data into intervals like distributions or probabilities E.g. Median, Quartiles, Percentiles.

Q-Q plots are important tools to validate the assumptions of linear regression like if the residuals from a model are indeed normally distributed or presence of the skewness.
While selecting the data for modelling, it can help assess if the data appears to be coming from the same population.
