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)

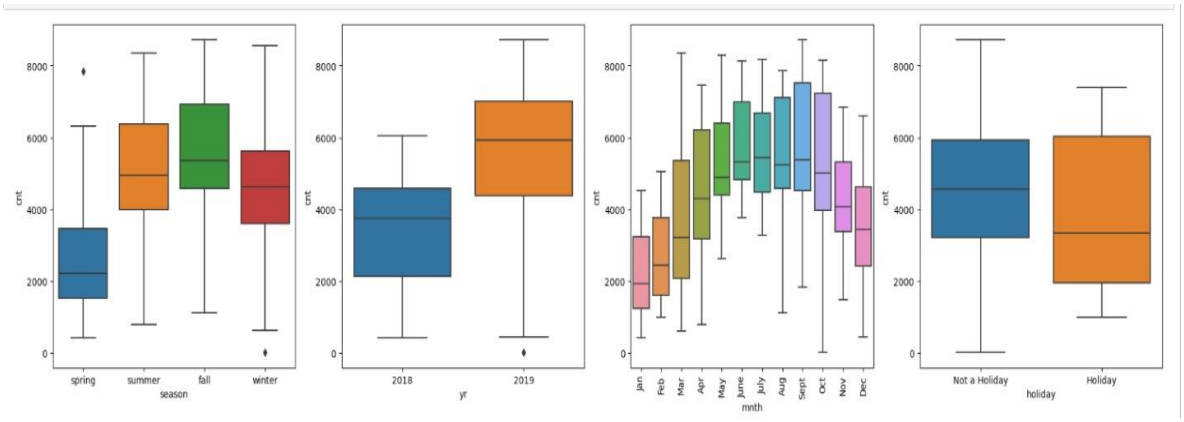
Total Marks: 3 marks (Do not edit)

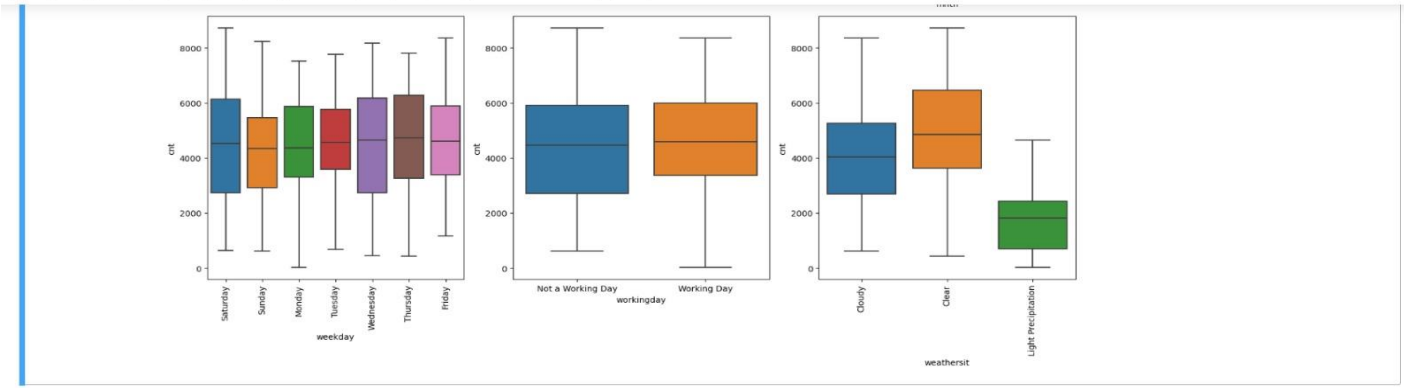
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Effect of categorical variables:

```
In [15]: #visualizing categorical variables
         plt.figure(figsize=(20,12))
         # 'season'
         plt.subplot(2,4,1)
         sns.boxplot(x='season',y='cnt',data=bike_sharing_df)
         #plt.title('Season')

         # 'yr'
         plt.subplot(2,4,2)
         sns.boxplot(x='yr',y='cnt',data=bike_sharing_df)
         #plt.title('Year')

         # 'mnth'
         plt.subplot(2,4,3)
         sns.boxplot(x='mnth',y='cnt',data=bike_sharing_df)
         plt.xticks(rotation=90)
         #plt.title('Months')

         #holiday
         plt.subplot(2,4,4)
         sns.boxplot(x='holiday',y='cnt',data=bike_sharing_df)
         #plt.title('Holiday')
```

```
         #weekday
         plt.subplot(2,4,5)
         sns.boxplot(x='weekday',y='cnt',data=bike_sharing_df)
         plt.xticks(rotation=90)
         #plt.title('Weekday')

         #workingday
         plt.subplot(2,4,6)
         sns.boxplot(x='workingday',y='cnt',data=bike_sharing_df)
         #plt.title('Working Day')

         #weathersit
         plt.subplot(2,4,7)
         sns.boxplot(x='weathersit',y='cnt',data=bike_sharing_df)
         plt.xticks(rotation=90)
         #plt.title('Weather Situation')

         plt.subplots_adjust(hspace=0.3,wspace=0.3)
         plt.tight_layout()
```





Observations & Insights of categorical variables

- Season(season): Among all four season, in fall season the bike sharing count is most followedby in summer as compare to other seasons.
- Year(yr): The median value of 2019 is highest and also the bike count in 2019 is maximum.
- Month(mnth) : If we see the trend of median values, generally in summer season i.e, july, august, september months, it is high. In July it is most. The bike sharing count is maximum in september month. Lowest bike rental count is for month 10 i.e. October.
- Median and highest bike rental count is for no holiday days marked with Not a Holiday.
- Bike rental median for all days are almost close to each other.
- Bike rental median for working/non-working days are almost close to each other.
- Bike rental median is highest for weathersit 2. Also maximum bike rental count is for weathersit 2 and lowest is for weathersit 3

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' when creating dummy variables helps prevent multicollinearity in your dataset. When you convert a categorical variable into dummy variables, each category becomes a separate binary column, with 1 indicating the presence of the category and 0 otherwise. Including all categories as dummy variables would result in redundancy: the information in one column can be inferred from theothers.

By setting 'drop_first=True', you drop one dummy variable (often called the "reference" category). Thisprevents multicollinearity by ensuring that the dummies aren't linearly dependent, which makes the regression model more stable and interpretable.

Example:
Without drop_first=True
Suppose we have a dataset with a categorical variable, Color, which has three categories: Red, Blue, and Green.
When we create dummy variables without drop_first=True, here's how the encoding looks:

Color	Color_Red	Color_Blue	Color_Green
Red	1	0	0
Blue	0	1	0
Green	0	0	1

this setup, if we know the values of any two dummy columns, we can deduce the value of the third.

For example:

If Color_Red and Color_Blue are both 0, we know the color must be Green.

This redundancy causes multicollinearity in regression, which makes the model's coefficient estimates unstable.

Example With drop_first=True

Using drop_first=True drops the first dummy variable (e.g., Color_Red). Here's what the encoding lookslike:

Color	Color_Blue	Color_Green
Red	0	0
Blue	1	0
Green	0	1

Now:

Red is represented by both Color_Blue and Color_Green being 0.

Blue and Green each have their own unique combination.

With this setup, there's no redundancy among the dummy variables, and we avoid multicollinearity.

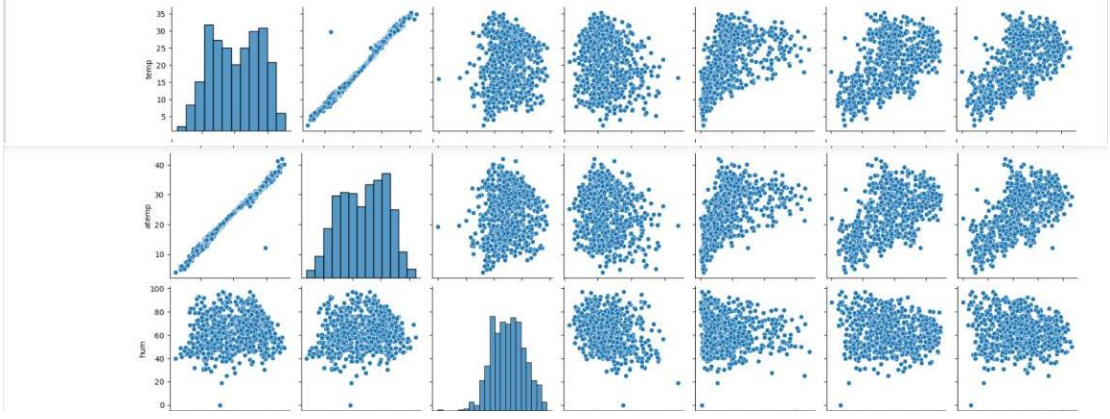
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

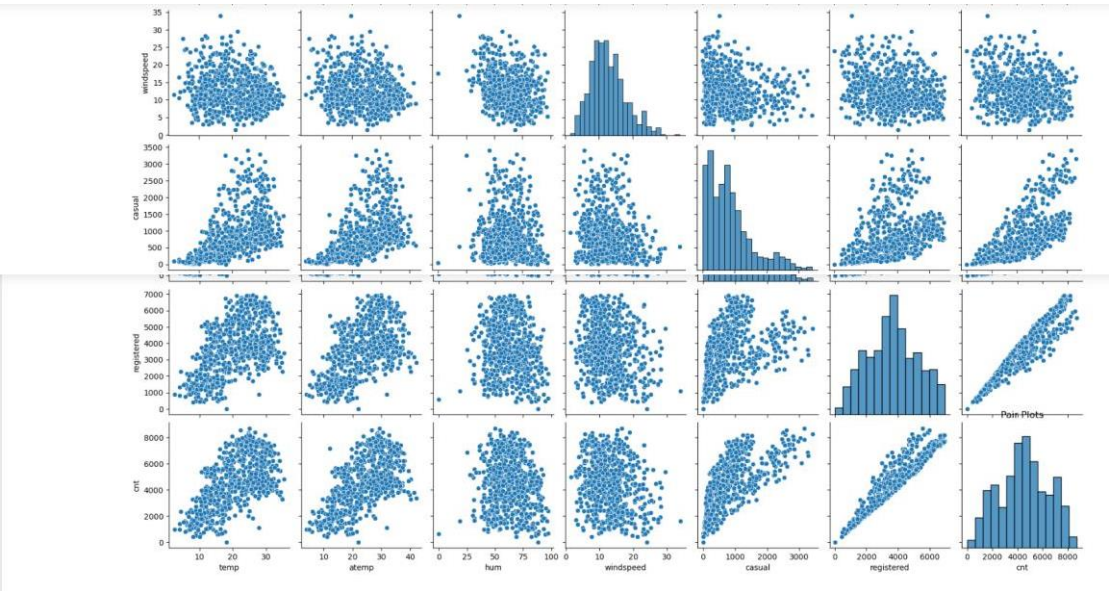
Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

```
In [13]: # Plot the pair-plot for numerical variables and analyse the plots
         plt.figure(figsize=(12,8))
         sns.pairplot(bike_sharing_df)
         plt.title("Pair Plots")
         plt.show()
```

```
<Figure size 1200x800 with 0 Axes>
```





Observations and Insights of above pair-plot:

- **Temperature(temp) and Adjusted Temperature(atemp) are highly correlated**
- Both temp and atemp show a positive correlation with cnt(total count), meaning higher temperatures tend to be associated with higher count of bike rentals.
- There is a strong linear relationship between temp and atemp, indicating potential multicollinearity if both variables are included in the regression model. It might be better toinclude only one of them to avoid redundancy.
- Humidity has a weaker, more scattered relationship with cnt, although there appears to be a slight negative trend. This suggests that higher humidity might slightly reduce bike rentals butisn't a strong predictor.
- The scatter plot of windspeeed against cnt shows no strong linear relationship, it likely has alimited influence on bike rental counts.
- Casual and registered users both have strong positive linear relationships with cnt. This isexpected, as cnt is the total count of bike rentals, including both casual and registered.
- Including both casual and registered might lead to multicollinearity, as cnt is derived from these two variables. It may be useful to model them independently or only use cnt directly.

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)
Validating the assumptions of Linear Regression after building the model on the training set is a crucial step to ensure the model is reliable and the inferences drawn from it are valid. Here's how you can validate the assumptions:

1. Linearity of the Relationship

- Check: The relationship between the independent variables and the dependent variable should be linear.
- Validation:
  - o Plot the observed vs. predicted values (a scatter plot). The points should lie along a straight diagonal line.
  - o Plot residuals (errors) vs. predicted values. The residuals should be randomly scattered around zero, with no obvious patterns.

2.Independence of Errors

• Check: The residuals (errors) should be independent.

• Validation:

o Use the Durbin-Watson test to check for autocorrelation in the residuals. A value close to 2 indicates no autocorrelation.

o Alternatively, you can plot residuals over time or order and look for patterns. Randomly scattered residuals indicate independence.

3.Homoscedasticity (Constant Variance of Errors)

• Check: The variance of the residuals should remain constant across all levels of the independent variables.

• Validation:

o Plot the residuals vs. predicted values. The spread of residuals should be constant (i.e., no funnel or cone shape). If there's a funnel shape, it indicates heteroscedasticity.

o You can also perform the Breusch-Pagan test or White test to statistically test for heteroscedasticity.

4.Normality of Residuals

• Check: The residuals should be approximately normally distributed.

• Validation:

o Histogram of residuals: Plot a histogram of the residuals. It should resemble a bell-shaped curve.

o Q-Q plot: Plot a quantile-quantile (Q-Q) plot of the residuals. The points should fall approximately along the diagonal line.

o Shapiro-Wilk test: This statistical test checks for normality. A high p-value (> 0.05) indicates that the residuals are normally distributed.

5. No Multicollinearity

• Check: The independent variables should not be highly correlated with each other.

• Validation:

o Calculate the Variance Inflation Factor (VIF) for each independent variable. A VIF value > 10 indicates high multicollinearity.

o You can also look at the correlation matrix of the independent variables. High correlation coefficients (near ±1) suggests multicollinearity.

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

1. Year:
    - **coefficient: 0.2461**
    - This feature has highest t-value (27.027),indicating strong positive relationship with bike demand . As the year advances, bike demand increases significantly.
2. Weekday(Weekend i.e. saturday):
    - **Coefficient : 0.0664**
    - This is another factor leading to increase in the demand of bike sharing.

3. WeatherSituation(weathersit_3)
   - **Coefficient: -0.3202**
   - This feature has a significant negative coefficient, indicating that adverse weather condition lead to reduction in the bike sharing service.

```
Df Model:                          12
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.5348      0.016     33.009      0.000       0.503       0.567
yr               0.2461      0.009     27.027      0.000       0.228       0.264
workingday       0.0571      0.012      4.600      0.000       0.033       0.081
windspeed       -0.1926      0.028     -6.836      0.000      -0.248      -0.137
spring          -0.2376      0.014    -16.537      0.000      -0.266      -0.209
summer          -0.0385      0.013     -3.070      0.002      -0.063      -0.014
Dec             -0.1183      0.017     -6.880      0.000      -0.152      -0.085
Jan             -0.1232      0.020     -6.315      0.000      -0.162      -0.085
Nov             -0.1122      0.018     -6.376      0.000      -0.147      -0.078
Sept             0.0563      0.018      3.116      0.002       0.021       0.092
Saturday         0.0664      0.016      4.146      0.000       0.035       0.098
Cloudy          -0.0890      0.010     -9.183      0.000      -0.108      -0.070
Light Precipitation -0.3202  0.027    -11.699      0.000      -0.374      -0.266
================================================================================
```

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 a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find a linear relationship between the variables by fitting a line (in the case of one feature) or a hyperplane(in the case of multiple features) that best predicts the target variable.

Steps Involved:

1. Assumptions:
   a. Linearity: The relationship between the independent and dependent variables is linear.
   b. Independence: Observations are independent of each other.
   c. Homoscedasticity: The variance of residuals (errors) is constant.
   d. Normality: The residuals of the model should be normally distributed.
2. Model Representation:
   a. Simple Linear Regression:
   $$y = \beta_0 + \beta_1 x + \epsilon$$
   y = \beta_0 + \beta_1x + \epsilon
   b. Multiple Linear Regression: \(y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n + \epsilon \)

c. Here, y is the dependent variable, \(x_1, x_2, \dots, x_n \) are independent variables, \beta_0 is the intercept, \(\beta_1, \beta_2, \dots, \beta_n \) are coefficients,and \epsilon is the error term.

3. Fitting the Model:
   - The coefficients \(\beta_0, \beta_1, \dots, \beta_n \) are estimated using methods like Ordinary Least Squares (OLS), which minimizes the sum of squared residuals \sum(\hat{y} - y)^2 ,where \hat{y} is the predicted value.
4. Model Evaluation:
   - The performance of the model is evaluated using metrics such as R-squared, Adjusted R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Linear regression is widely used in predictive modeling, but its accuracy depends on the validity of the assumptions and the quality of the input data.

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(mean, variance, correlation, etc.) but appear very different when graphed. Created by statistician Francis Anscombe in 1973, the quartet illustrates the importance of graphing data before analyzing it and shows how different datasets can have similar statistical properties but very different distributions.

Datasets:
- The quartet consists of four sets of (x, y) pairs.
- All four datasets have the same mean, variance, and correlation for x and y, and thesame regression line.

Importance:
- The first dataset shows a linear relationship that fits well with linear regression.
- The second dataset has a clear non-linear relationship, suggesting that linearregression is not appropriate.
- The third dataset is a linear relationship with an outlier, which significantly affectsthe regression results.
- The fourth dataset shows a vertical line where one point is an outlier in the xdirection, resulting in amisleading regression line.

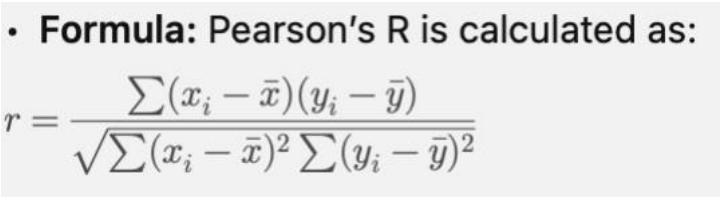
Conclusion:
Anscombe's quartet highlights the importance of visualizing data to identify patterns, outliers, and the appropriate statistical models for analysis.

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>

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between them.

- Formula: Pearson's R is calculated as:



r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} where x_i and y_i are the individual data points, and \bar{x} and \bar{y} are the means of the variables.

- Range: The value of Pearson's R ranges from -1 to 1.
- r = 1 indicates a perfect positive linear relationship.
- r = -1 indicates a perfect negative linear relationship.
- r = 0 indicates no linear relationship.

Importance:

Pearson's R is widely used in statistics to understand the strength and direction of the linear relationship between two variables. However, it only measures linear relationships and can bemisleading if the relationship is non-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 the process of adjusting the range of independent variables or features of data so that they fit within a specific scale, such as 0-1 or have a mean of 0 and a standard deviation of 1. Scaling is important in algorithms that compute distances between data points, such as k-nearest neighbors or support vector machines, or when gradient descent optimization is used, as in linear regression.

Types of Scaling:

1. Normalized Scaling:
   - Definition: Normalization scales the features to a fixed range, typically 0 to 1.
   - Formula :

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

     Formula: x{\prime} = \frac{x - \min(x)}{\max(x) - \min(x)}
   - Use Case: Useful when the features have different units and you want to bring themto a common scale.
2. Standardized Scaling:
   - Definition: Standardization scales the features such that they have a mean of 0 and a standard deviation of 1.
   - Formula: x{\prime} = \frac{x - \mu}{\sigma}
   - Use Case: Useful when the features have different units and distributions but youwant them to have the same statistical properties.

Difference:

- Normalization is bound to a specific range, which is particularly useful when you need to ensure that the features have the same scale.
- Standardization adjusts the data to a normal distribution with a mean of 0 and a standard deviation of 1, which is useful for algorithms that assume normally distributed data.

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>

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression models. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

Infinite VIF:

- VIF becomes infinite when there is perfect multicollinearity between the features,meaning one predictor variable is a perfect linear combination of other predictors.
- In other words, if a feature can be exactly predicted by one or more other features inthe dataset, the denominator in the VIF calculation, which is 1 - R^2 (where R^2 is the coefficient of determination), becomes zero, leading to an infinite VIF value.Implications:
- An infinite VIF indicates that the regression model has redundant features, making itunstable. Removing or combining the collinear features can help stabilize the model.

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 (Quantile-Quantile Plot) is a graphical tool used to assess whether a dataset follows a particular distribution, most commonly the normal distribution. It plots the quantiles of the dataset against the quantiles of a theoretical distribution (like the normal distribution).

How to Interpret:

- If the points in a Q-Q plot fall along a straight line, it suggests that the data followsthe theoretical distribution.
- Deviations from the straight line indicate departures from the theoreticaldistribution.

Use in Linear Regression:

- In linear regression, a key assumption is that the residuals (errors) are normallydistributed. A Q-Q plot of the residuals can be used to check this assumption.
- If the residuals follow a straight line in the Q-Q plot, it confirms that they areapproximately normally distributed, validating the assumption.
- If the residuals deviate significantly from the straight line, it suggests that thenormality assumption may be violated, potentially affecting the validity of the regression model.
- Importance:
- The Q-Q plot is crucial for diagnosing issues with model assumptions, helping toensure that the linear regression model is appropriate for the data at hand.