# 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)**

    - Season Fall has the highest median, which makes absolute sense as weather is beautiful in fall and people would ride more bikes.
    - Year 2019 has higher median, which could hind towards bike riding becoming more popular as compared to 2018.
    - As expected during winter months, the median is low. And during the months of Summer and Fall, the median is higher indicating that people prefer more bike riding.
    - On Sunday, the median is low, indicating that people prefer to stay home or spend time with families and not go out biking.
    - As expected in clear weather, the median is higher as people prefer not to bike when weather is not clear.
    - As expected on a holiday the median is lower, indicating low demand for biking.

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

    When creating dummy variables for use in a regression model, the drop_first=True parameter is typically the one that should be used. This is due to the fact that dummy variables in a regression model are typically collinear, or highly correlated. Collinearity can make the model more susceptible to overfitting and make it difficult to interpret the regression coefficients.

    In a regression model, categorical data are encoded with dummy variables. We could, for instance, use three dummy variables to encode a column with three possible values (A, B, or C) in a dataset:

    For rows with a value of A in the original column, D1 = 0, D2 = 0, and D3 = 1

    For rows with a value of B in the original column, D1 = 0, D2 = 1, and D3 = 0

    For rows with a value of C in the original column, D1 = 1, D2 = 0, and D3 = 0

Because the regression coefficients for the dummy variables won't be uniquely defined, collinearity can cause issues in a regression model. Because of this, we will not be able to interpret the coefficients in the usual manner, and the regression results will not be reliable.

This issue is resolved by eliminating one of the dummy variables for each categorical column with the drop_first=True parameter. In the previous scenario, for instance, we would eliminate D1 from our model, limiting it to only D2 and D3. As a result, the regression coefficients will be uniquely defined and the dummy variables will no longer be collinear. This improves the model's reliability and makes it possible for us to interpret the coefficients in a meaningful manner. In conclusion, it is essential to use the drop_first=True parameter when creating dummy variables to prevent collinearity in the regression model and to meaningfully interpret the regression coefficients.

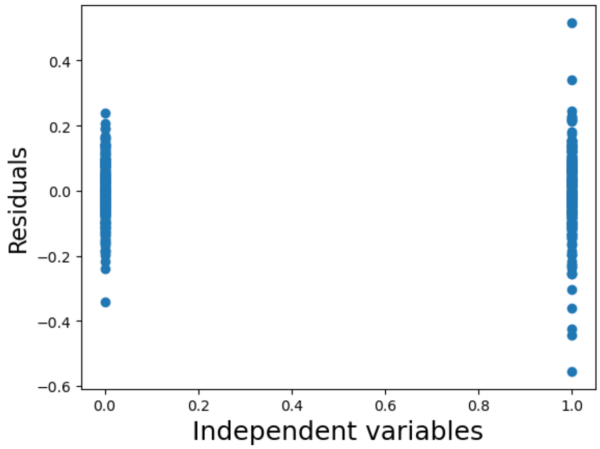
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Variables ‘atemp’ and ‘temp’ have the highest correlation with the target variable.

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

   - Assumptions of homoscedasticity and correlation are validated by below plot performed after the model has been created

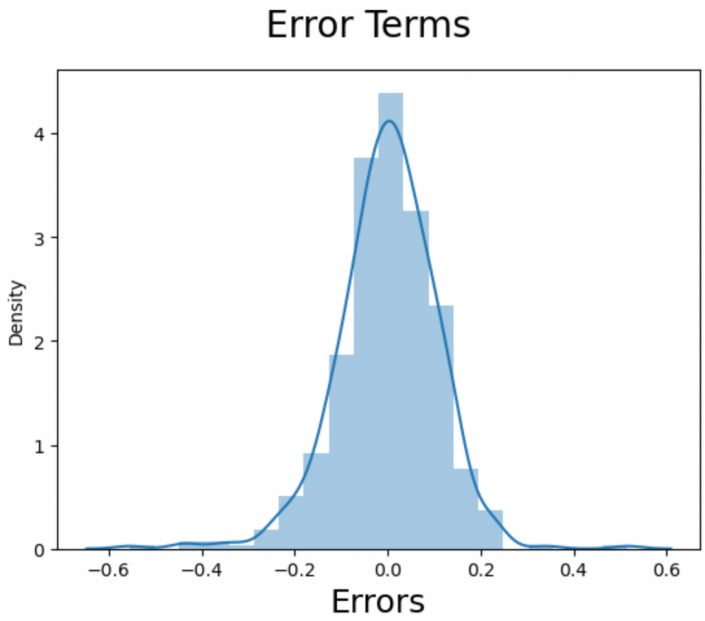
```
In [198]: #PLotting the residuals to see if a pattern exists
          #Checking assumption of homoscedasticity and autocorrelation
          plt.figure()
          plt.scatter(X_t,res)
          fig.suptitle('Independent vars vs res', fontsize=20)              # Plot heading
          plt.xlabel('Independent variables', fontsize=18)                  # X-label
          plt.ylabel('Residuals', fontsize=16)
          plt.show()
```




   - Assumption of error terms normally distributed with mean 0 are validated by below plot performed after the model is created

```
In [184]: #Checking ASSUMPTION OF NORMALITY:
          # Plot the histogram of the error terms
          fig = plt.figure()
          sns.distplot((res), bins = 20)
          fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
          plt.xlabel('Errors', fontsize = 18)                         # X-label

Out[184]: Text(0.5, 0, 'Errors')
```




- Assumption of linearity between X and Y is explained and validated by the R squared score of 0.77, which validates that 77% of the variance in the data is explained by the model.

**5. Based on the final model, which are the top 3 features contributing significantly towardsexplaining the demand of the shared bikes?**

Based on the final model, below are the coefficients for various variables on which the dependent variable 'cnt' depends:

| Variable | Coefficient Value |
|---|---|
| const | 0.4687 |
| yr | 0.2466 |
| Month 9 | 0.1465 |
| Month 10 | 0.1002 |
| Month 8 | 0.0963 |
| Month 7 | 0.0678 |
| Month 5 | 0.0661 |
| winter | -0.0393 |
| Sunday | -0.0512 |
| Mist cloudy | -0.0972 |
| holiday | -0.1031 |

| | |
|---|---|
| spring | -0.2383 |
| Light rain Light snow Thunderstorm | -0.3307 |

Based on the final model selected, the equation for the best fitted line is :

cnt=0.2466 * yr + 0.0661*5 + 0.09663*8 + 0.1465*9 + 0.0678*7 + 0.1002*10 - 0.1031*holiday -0.2383*spring - 0.0393*winter - 0.3307* light rain light snow thunderstorm - 0.0972* Mist Cloudy - 0.0512* Sunday + 0.4687

As evident from the coefficients, the demand will increase the highest when there's an increase in the value of variables 'yr' and 'month 9'. And demand will decrease the most when there's increase in 'Light rain Light snow Thunderstorm' or the season is 'spring' and if it's a 'holiday'.

Recommendations for the company:

Demand will be higher in the months numbered 9, 10, 8,7,5 and the demand will be lowest when weather is light rainlight snow thunderstorm, when the season is spring or winter , when its a holiday or a sunday and weather is misty cold.

And the top 3 features contributing to the demand are –

1. which month is it ? (e.g. demand highest in month 9)

2. how's the weather and which season is it? (e.g. demand lowest in winter and spring or when its cold and misty and when there's thunderstorm)

3. if it's a holiday. (e.g. on Sundays demand is low)

# General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes that the relationship between the dependent and independent variables is linear, which means that a change in the dependent variable is directly proportional to the change in the independent variable(s).

Linear regression algorithm is a supervised learning algorithm, which requires a labeled dataset that includes both the input features and the corresponding correct output values. Algorithm finds the line of best fit, which is a line that represents the relationship between the independent and dependent variables in the dataset as accurately as possible.

To find the line of best fit, the algorithm starts by initializing the coefficients of the line to some random values. Then, it iteratively improves these coefficients by using a optimization algorithm, such as gradient descent.At each iteration, the algorithm makes a prediction using the current values of the coefficients, and then it compares this prediction to the correct output value for the input features. Based on this comparison, it calculates an error value, which represents how far off the prediction was from the correct output.

The algorithm then uses this error value to update the coefficients of the line in a way that will reduce the error in future predictions. This process continues until the algorithm reaches a pre-determined stopping criteria, such as a maximum number of iterations or a minimum error threshold.Once the optimization process is complete, the algorithm outputs the final values of the coefficients as the line of best fit. This line can then be used to make predictions on new input data, by plugging the input values into the equation of the line and solving for the predicted output value.

Overall, the linear regression algorithm is a relatively simple and a way to model the relationship between two or more variables in a dataset. It is widely used in a variety of applications, such as predicting stock prices, analyzing consumer behavior, and understanding the relationship between various health and environmental factors.

There are two types of linear regressions: Simple and Multiple. Simple linear regression is used to model relationship between one dependent and one independent variable. Wherea

Multiple linear regression is used to model relation between multiple independent and one dependent variable.

**2. Explain the Anscombe’s quartet in detail.**

Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. The four datasets have very similar statistical properties, such as mean, variance, and correlation, but they have very different visualizations when plotted on a graph. This demonstrates the importance of visualizing data before performing statistical analyses, as relying on summary statistics alone can sometimes be misleading.

Each of the four datasets in Anscombe's quartet consists of 11 (x, y) pairs, for a total of 44 data points. The first dataset has a linear relationship between x and y, the second has a quadratic relationship, the third has a relationship that can be modeled by a rectangular hyperbola, and the fourth has a relationship that is essentially random. However, despite these differences, the statistical properties of the four datasets are almost identical.

For example, the mean of x and y are both the same for all four datasets (approximately 9), the variance of x and y are also the same for all four datasets (approximately 11), and the correlation between x and y are also the same for all four datasets (approximately 0.816).

Despite the statistical similarities, when these datasets are visualized on a scatter plot, it becomes clear that they have very different patterns and structures. For example, the first dataset has a strong linear relationship between x and y, while the second dataset has a more complex quadratic relationship. The third dataset has a relationship that is roughly linear near the center of the plot, but becomes more dispersed as the values of x and y increase. And the fourth dataset has no clear relationship at all, with the points scattered randomly across the plot.

3. **What is Pearson's R?**

Pearson's R is a measure of the linear correlation between two variables. It ranges from -1 to 1, where a value of -1 indicates a perfect negative correlation (i.e., as one variable increases, the other variable decreases), a value of 0 indicates no correlation, and a value of 1 indicates a perfect positive correlation (i.e., as one variable increases, the other variable also increases). It is commonly used to determine the strength of a linear relationship between two variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling**

   **and standardized scaling?**

   Scaling is the process of transforming data points in a dataset to fall within a specific range, such as 0 to 1 or -1 to 1. This is typically performed to improve the performance of a machine learning model, since many algorithms perform better when the data is within a certain range.

   There are two main types of scaling: normalized scaling and standardized scaling.

   Normalized scaling, also known as min-max scaling, scales the data such that the minimum value of the data becomes 0 and the maximum value becomes 1. This is done by subtracting the minimum value from each data point and dividing by the range of the data (i.e., the maximum value minus the minimum value).

   Standardized scaling, also known as z-score scaling, scales the data such that the mean of the data is 0 and the standard deviation is 1. This is done by subtracting the mean of the data from each data point and dividing by the standard deviation of the data.

   Both normalized scaling and standardized scaling are commonly used to improve the performance of machine learning algorithms, but standardized scaling is generally more useful when the data has a Gaussian distribution (i.e., a bell-shaped curve).

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

The value of the variance inflation factor (VIF) can be infinite if there is a perfect linear relationship between two or more variables in a multiple regression model. This means that one of the variables can be perfectly predicted from the others, so there is no variance left to explain. In this case, the VIF is infinite because the denominator of the VIF formula is 0 (i.e., the variance of the predicted values is 0).

Having a VIF of infinite indicates that the model is not reliable, since the predictions made by the model will be overly certain and the coefficients will be poorly estimated. To fix this, one or more of the variables with a VIF of infinite must be removed from the model.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to check whether a given set of data follows a specific distribution. It plots the quantiles of the data against the quantiles of a theoretical distribution, such as the normal distribution.

In linear regression, a Q-Q plot is often used to check whether the residuals (i.e., the differences between the observed values and the predicted values) follow a normal distribution. This is important because many statistical tests and confidence intervals used in linear regression rely on the assumption that the residuals are normally distributed.

If the Q-Q plot shows that the residuals do not follow a normal distribution, it may indicate that the linear regression model is not appropriate for the data, or that there are significant outliers in the data. In either case, further investigation may be needed to improve the model or the data.