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)

   Ans:

   - *Season: 3: fall has highest demand for rental bikes.*
   - *The demand for next year has grown.*
   - *Demand is continuously growing each month till June. September month has highest demand. After September, demand decreases.*
   - *When there is a holiday, demand decreases.*
   - *The 'good' weathersit has highest demand.*
   - *During September, bike sharing is more. During the year end and beginning, demand is less compared to rest of the year.*

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

   Ans:

   - *drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.*
   - *Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.*
   - *Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.*

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

   Ans:

   - 'temp' and 'atemp' both variables together have the highest correlation with the target variable.

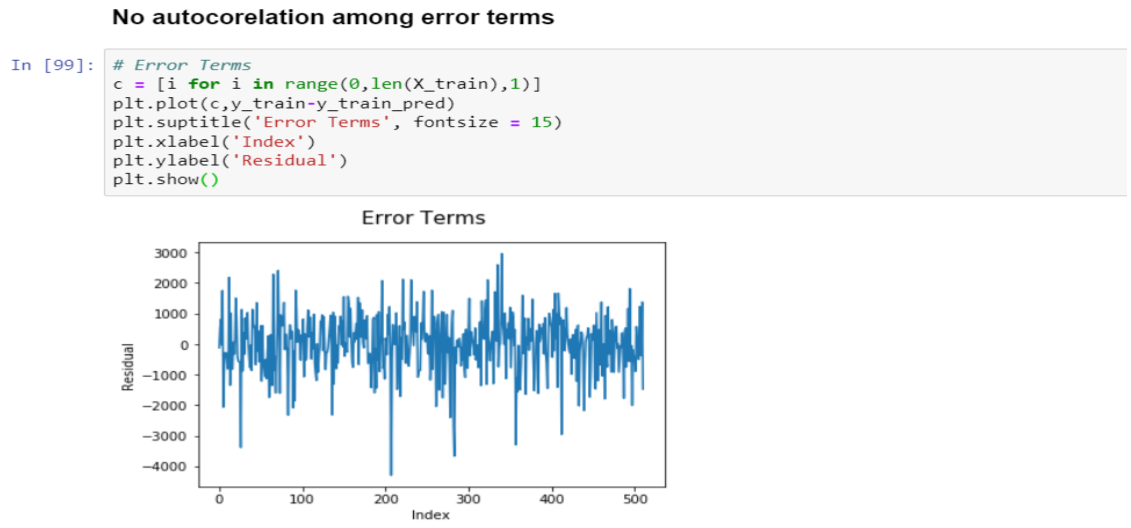
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Ans:

Following are the assumptions that needed to be validated:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other.
- Error terms have constant variance (homoscedasticity)

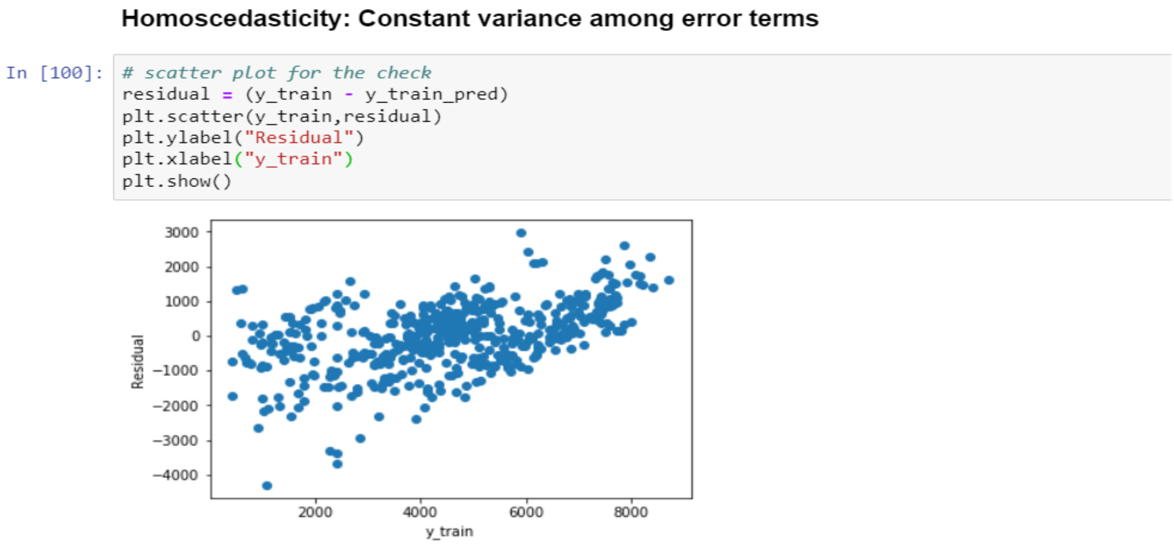
Following are the steps followed to do it:



No autocorelation among error terms

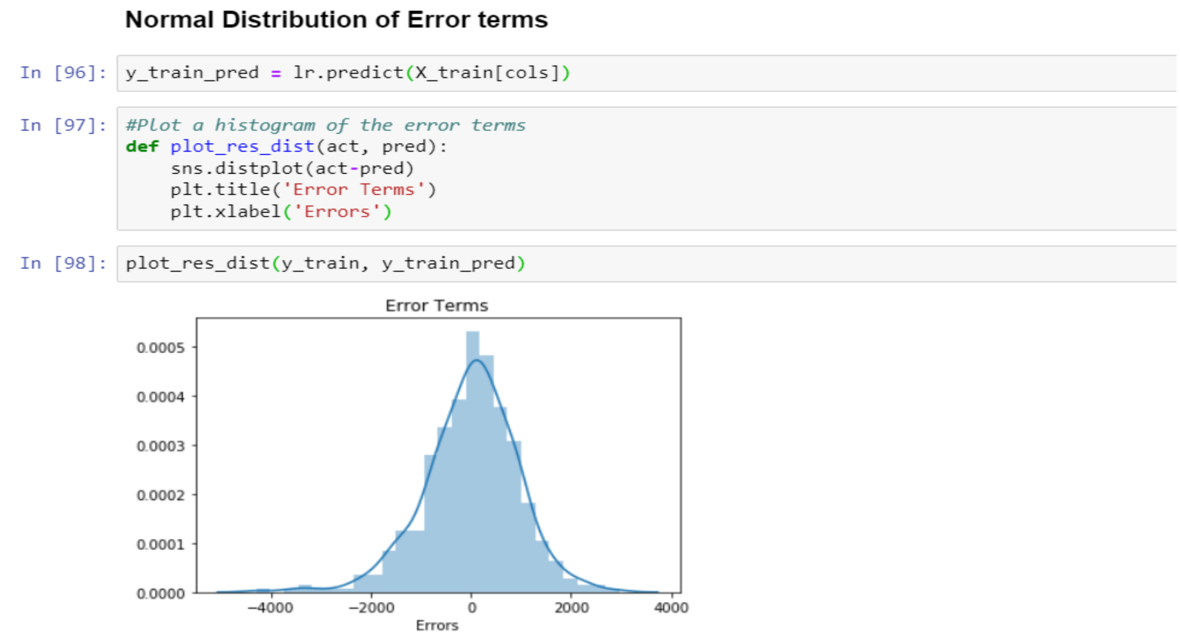
In [99]:
```
# Error Terms
c = [i for i in range(0,len(X_train),1)]
plt.plot(c,y_train-y_train_pred)
plt.suptitle('Error Terms', fontsize = 15)
plt.xlabel('Index')
plt.ylabel('Residual')
plt.show()
```

Error Terms



Homoscedasticity: Constant variance among error terms

In [100]:
```
# scatter plot for the check
residual = (y_train - y_train_pred)
plt.scatter(y_train,residual)
plt.ylabel("Residual")
plt.xlabel("y_train")
plt.show()
```

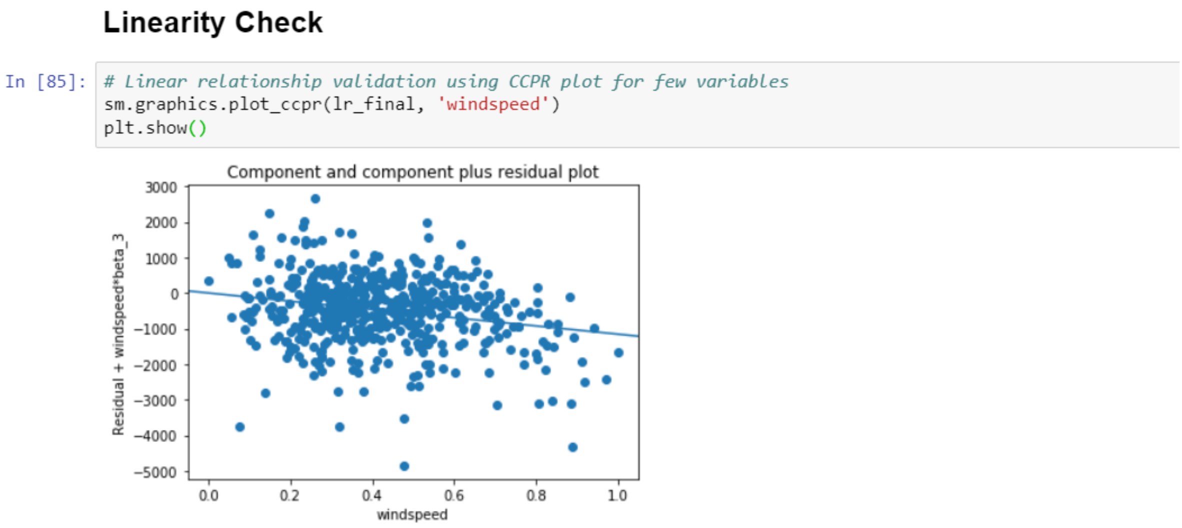


Normal Distribution of Error terms

```
In [96]: y_train_pred = lr.predict(X_train[cols])
```

```
In [97]: #Plot a histogram of the error terms
         def plot_res_dist(act, pred):
             sns.distplot(act-pred)
             plt.title('Error Terms')
             plt.xlabel('Errors')
```

```
In [98]: plot_res_dist(y_train, y_train_pred)
```



Linearity Check

```
In [85]: # Linear relationship validation using CCPR plot for few variables
         sm.graphics.plot_ccpr(lr_final, 'windspeed')
         plt.show()
```

Component and component plus residual plot
Residual + windspeed*beta_3
3000
2000
1000
0
−1000
−2000
−3000
−4000
−5000
0.0
0.2
0.4
0.6
0.8
1.0
windspeed

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

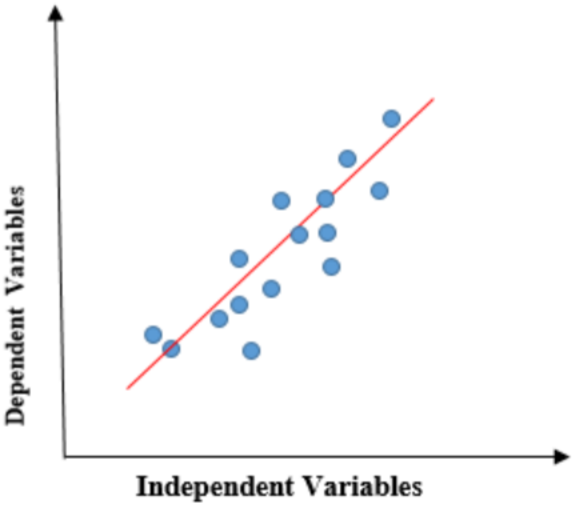
Following are the top 3 predictors as they have the largest coefficients.

- Temp
- Weathersit (bad)
- Year

General Subjective Questions

6. Explain the linear regression algorithm in detail. (4 marks)

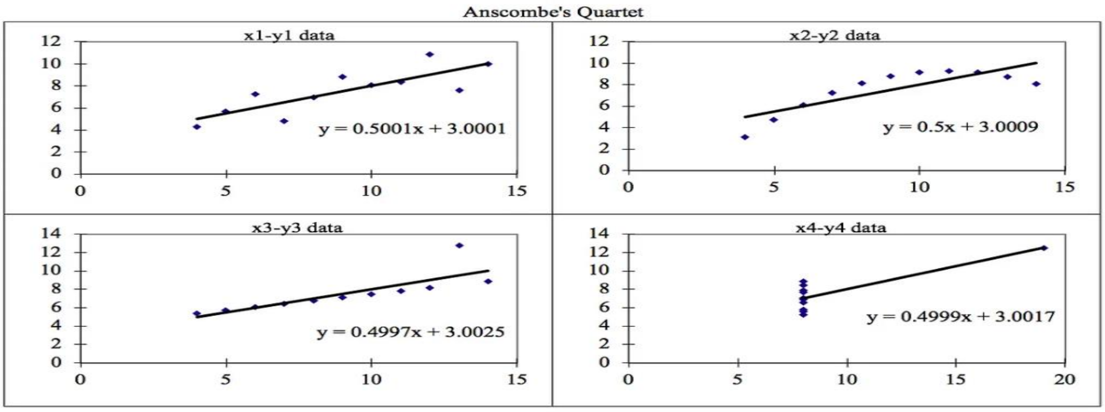
Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. *If there is a single input variable (x), such linear regression is called* ***simple linear regression****. And if there is more than one input variable, such linear regression is called* ***multiple linear regression*****.** The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

7. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.



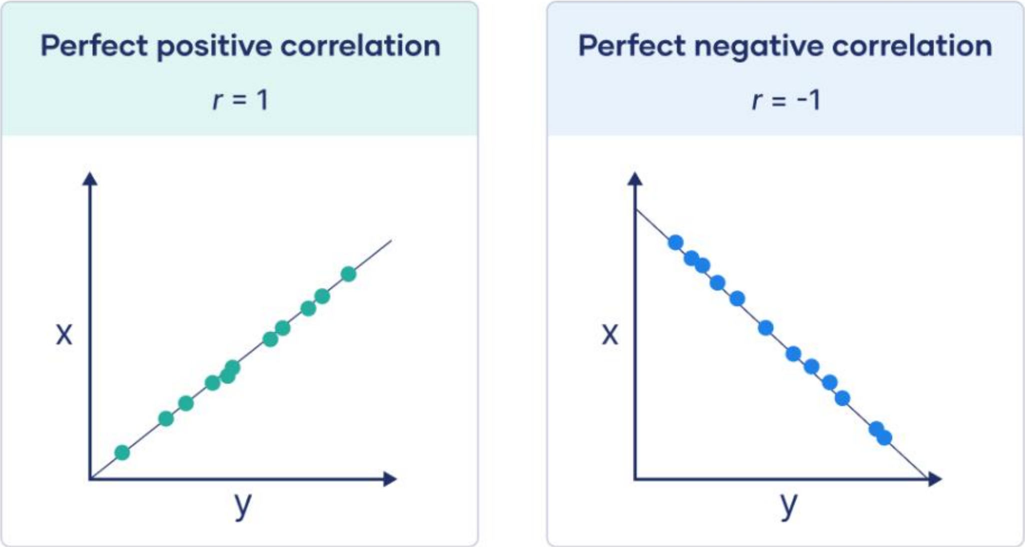
8. What is Pearson’s R? (3 marks)

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Another way to think of the Pearson correlation coefficient (*r*) is as a measure of how close the observations are to a line of best fit.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.

When *r* is 1 or –1, all the points fall exactly on the line of best fit.



9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature 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.

In machine learning, feature scaling is employed for a number of purposes:

- Scaling guarantees that all features are on a comparable scale and have comparable ranges. This process is known as feature normalisation. This is significant because the magnitude of the features has an impact on many machine learning techniques. Larger scale features may dominate the learning process and have an excessive impact on the outcomes. You can avoid this problem and make sure that each feature contributes equally to the learning process by scaling the features.
- Algorithm performance improvement: When the features are scaled, several machine learning methods, including gradient descent-based algorithms, distance-based algorithms (such k-nearest neighbours), and support vector machines, perform better or converge more quickly. The algorithm’s performance can be enhanced by scaling the features, which can hasten the convergence of the algorithm to the ideal outcome.
- Preventing numerical instability: Numerical instability can be prevented by avoiding significant scale disparities between features. Examples include distance calculations or

matrix operations, where having features with radically differing scales can result in numerical overflow or underflow problems. Stable computations are ensured and these issues are mitigated by scaling the features.

- Scaling features makes ensuring that each characteristic is given the same consideration during the learning process. Without scaling, bigger scale features could dominate the learning, producing skewed outcomes. This bias is removed through scaling, which also guarantees that each feature contributes fairly to model predictions.

Normalization	**Standardization**
This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
When features are on various scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

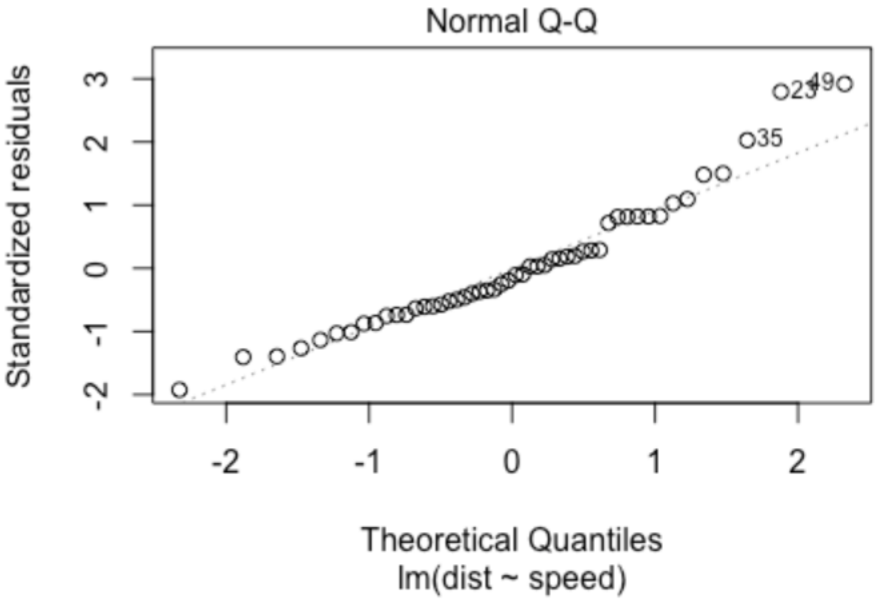
The formula of VIF is given by:
VIF = 1/(1-R^2) [where R -> Coefficient of Determination]

If there is perfect correlation, then R^2 = 1, which will cause the VIF to be infinity. This means there is a variable which can explain the independent variable all by itself. There are no error terms.

11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

QQ plots in Linear Regression:
QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small sample sizes, you can't assume your estimator $\hat{\beta}$ is Gaussian either, so the standard confidence intervals and significance tests are invalid. However, it's worth trying to understand how the plot is created in order to characterize observed violations.



The points approximately fall on the line, but what does this mean? The simplest explanation is as follows: say you have some observations and you want to check if they come from a normal distribution. You can standardize them (mean center and scale variance to 1) and then ‘percentile match’ against a standard normal distribution. Then you can plot your points against a perfectly percentile-matched line.