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?

From our analysis of the categorical variables from the dataset, we can infer below points

- **The demand of bikes is less in the spring and high in the fall season when compared with other seasons**.
- **The demand of bike increased in the year 2019 when compared with year 2018**.
- **The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.**
- **The month box plots indicates that more bikes are rent during September month.**
- **The weekday box plots indicates that more bikes are rent during Saturday.**
- **The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, partly cloudy weather.**

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

- **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.

- Example: From our assignment, we have 4 types of values in Categorical column "**season**" and we want to create dummy variable for that column. If one variable is not Spring, Summer or winter then It is obvious Fall. So we do not need 4th variable to identify the Fall

Seasons	Spring	Summer	winter
Spring	1	0	0
Summer	0	1	0
winter	0	0	1
Fall	0	0	0

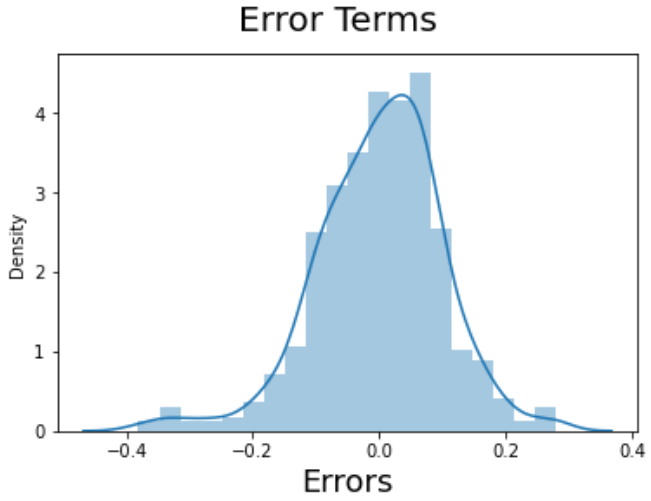
3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- By looking at the pair plot variables “temp” and “atemp” has the highest (0.63) correlation with target variable 'cnt'.

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

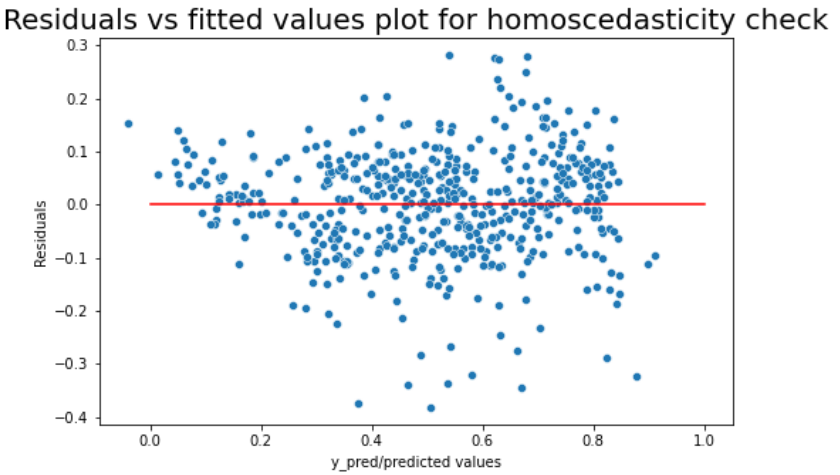
- Independent variables and the dependent variables could be transformed so that the relationship between them is linear.
- To validate the assumptions 2 methods are followed.

Residual analysis



From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid

Check for Homoscedasticity



From the above plot, we can see that residuals have equal or almost equal variance across the regression line.

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

The Top 3 features contributing significantly towards the demands of share bikes are:

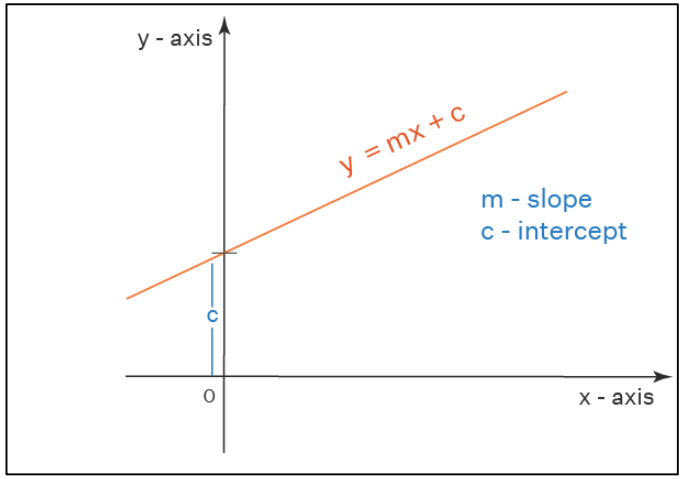
- aTemperature(Positive correlation)
- Year 2019 (Positive correlation)
- Weathersit Light snow and Rain (Negative correlation)

General Subjective Questions:

1) Explain the linear regression algorithm in detail.

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.
- It is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.
- Mathematically we can write Linear regression equation as

- y=mx+c
- Where m = Slope of the line
- c = y-intercept of the line
- x = Independent variable from dataset
- y = Dependent variable from dataset



Note: Image courtesy Google

1) Explain the linear regression algorithm in detail.

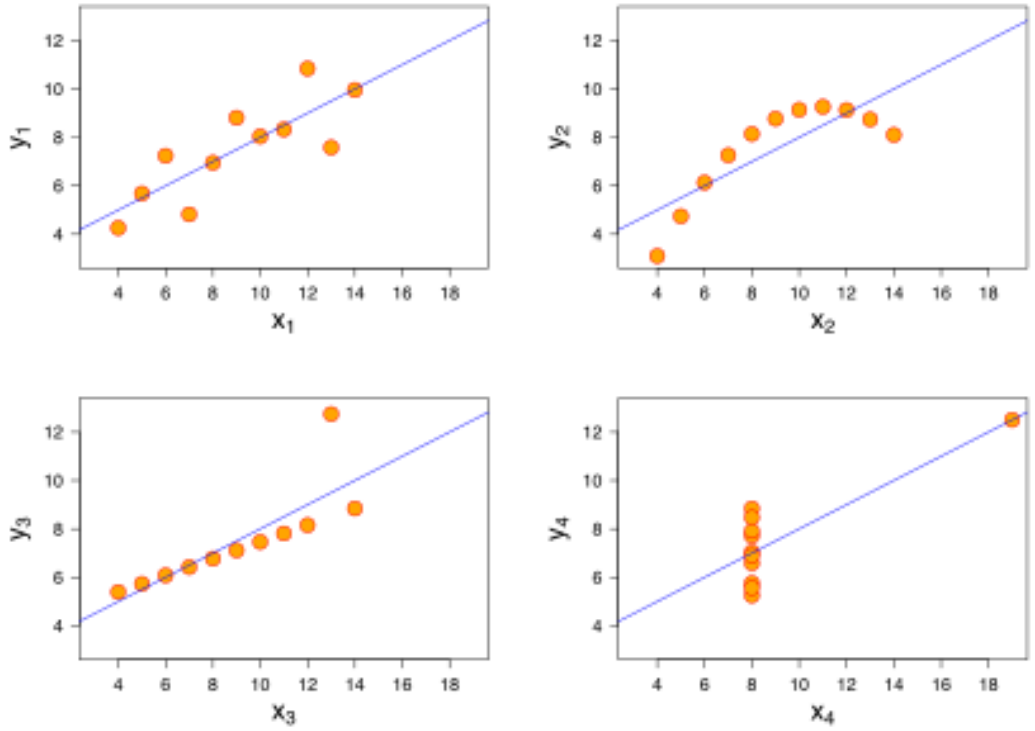
- Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.
- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

2) Explain the Anscombe's quartet in detail.

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 before analyzing it and the effect of outliers on statistical properties.
- Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points
- After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

2) Explain the Anscombe's quartet in detail.

- All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



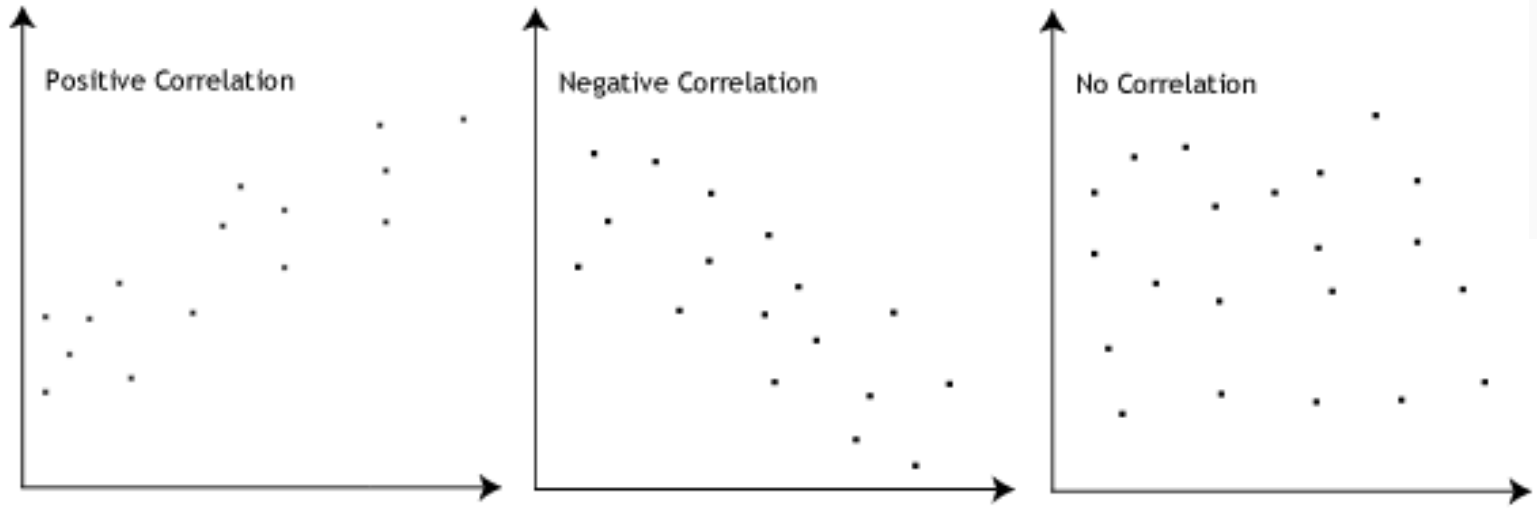
Note: Image courtesy Google

3) What is Pearson’s R?

- In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by 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.
- The Pearson's correlation coefficient varies between -1 and +1 where:
- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

3) What is Pearson's R?

- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r=correlation coefficient
- x_i=values of the x-variable in a sample
- $\bar{x}$=mean of the values of the x-variable
- y_i=values of the y-variable in a sample
- $\bar{y}$=mean of the values of the y-variable

Note: Image courtesy Google

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- **Scaling:** It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- **Why scaling:** Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- **Normalized scaling vs standardized scaling** :

Normalisation	Standardisation
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

- It so happens that sometimes some variables are able to create perfect multiple regressions on other variables
- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R^2 =1, which lead to 1/(1-R^2) infinity.
- To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multi collinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.
- If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

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

- Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.