Assignment-based Subjective Questions

Q. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Here are some of the inferences that I could make regarding the effect of independent categorical variables to our dependent variable "cnt":

- Fall season from the months "Aug" till "Oct" has attracted significantly high number of bookings. Also, the number bookings in each season has shown a significant rise from the year 2018 to 2019.
- The number of booking has shown an increasing trend from the starting of the year however, once we approach the end of the year, especially after "Oct" month, it shows a decreasing trend.
- Again, it does infer one more thing that a clear weather attracts more booking in general.
- Once we approach the weekend, especially from Thursday till Sunday, the number of booking is comparatively high.
- Also, it seems people book less during holidays since, they may want to spend time at home with their family rather than going out.
- However, booking doesn't seem to have any significant dependency on the factor whether it's working day or not.

Q. Why is it important to use drop_first=True during dummy variable creation?

Answer: drop_first=True helps in reducing an extra unwanted column during the dummy variable creation since, with or without it, we will infer the same information.

Also, with the removal of that extra column, it does reduce our effort to manually remove independent correlated variables (like we do using VIF).

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

Answer: 'temp' variable has the highest correlation with the target variable 'cnt'.

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

Answer: I have validated the assumptions of my Linear regression model based on the following factors below:

- **Normality of Error Terms:** The distribution of the error terms should be a normal distribution.
- **Multi-collinearity Check:** There should be minimal collinearity between the independent variables which was taken care in the assumption I made.
- **Homoscedasticity:** There should be no visible pattern in the scatter plot distribution of the residual values which was again taken care.

- **Independence of Residuals:** Ensured that there is no auto-correlation. Durbin-Watson's value of final model lm6 is 2.085, which signifies there is no first-order autocorrelation since the value lies very near to 2.00.

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

Answer: If we keep the "Year" variable aside, the top the features which explains the demand of the variable comes out to be:

- **temp**
- **Weather having Light_snowrain (Note: this was a category assigned by me on weathersit column)**
- **Holiday and winter season**

Note: Here I have taken those variables in a count which has a negative correlation, since that also counts under explaining the trend of our dependent variable.

General Subjective Questions

Q. Explain the linear regression algorithm in detail.

Answer: Linear Regression is **a machine learning algorithm based on supervised learning**. It is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. 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 a continuous numerical variable.

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 a linear regression equation as:

$$y = a + bx$$

Or, it can also be written in the form of hypothesis function for linear regression as: y= a1 + a2.x

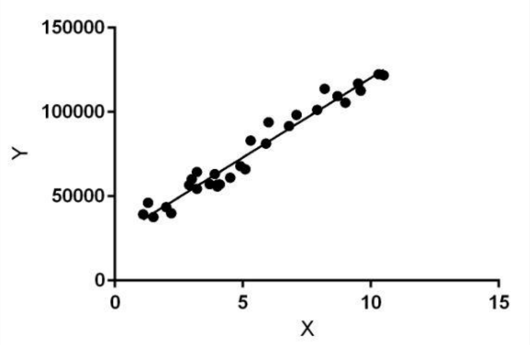
Where,

b/a2 = Slope of the line

a/a1 = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset



While training the model we are given:
x: input training data (univariate – one input variable (parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
a1: intercept
a2: coefficient of x
Once we find the best a1 and a2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y). We make use of **Cost Function** for the same.

Cost function of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

We need to update a1 and a2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent.

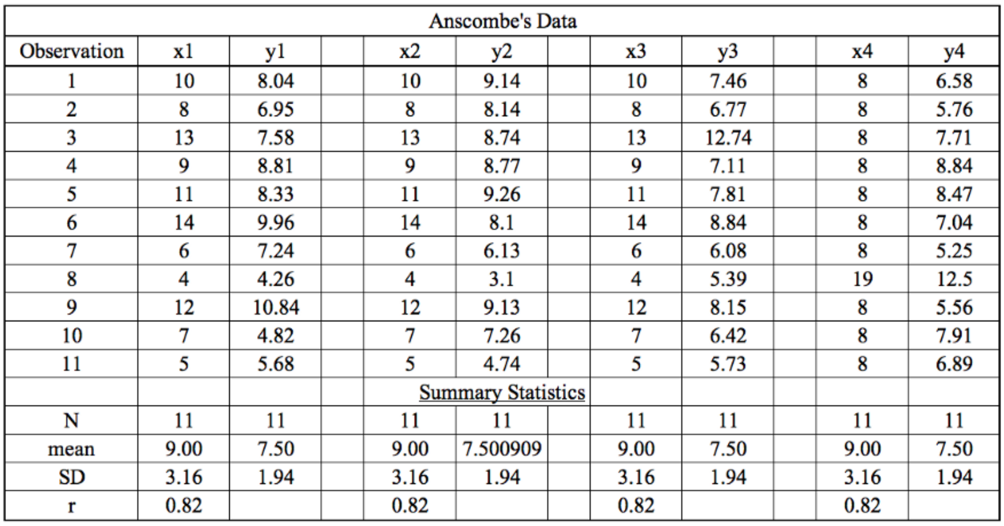
There are certain assumptions that we are supposed to make while doing the linear regression:

- **Normality of Error Terms:** The distribution of the error terms should be a normal distribution.
- **Multi-collinearity Check:** There should be minimal collinearity between the independent variables.
- **Linearity check:** We need to assume the linear relationship between the independent and dependent variables.
- **Homoscedasticity:** There should be no visible pattern in the scatter plot distribution of the residual values.
- **Independence of Residuals:** We need to ensure that there is no auto-correlation which can be done by checking **Durbin-Watson's** value of final model which if lies around the value of 2.00 then, there is no auto correlation.

Q. Explain the Anscombe’s quartet in detail.

Answer: They were developed by the statistician Francis Anscombe to demonstrate the importance of graphing data before analysing it and the effect of outliers on statistical properties. Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each of the 4 datasets consist of eleven data points (x, y).

The statistical information for all these four datasets are approximately similar and can be computed as follows:

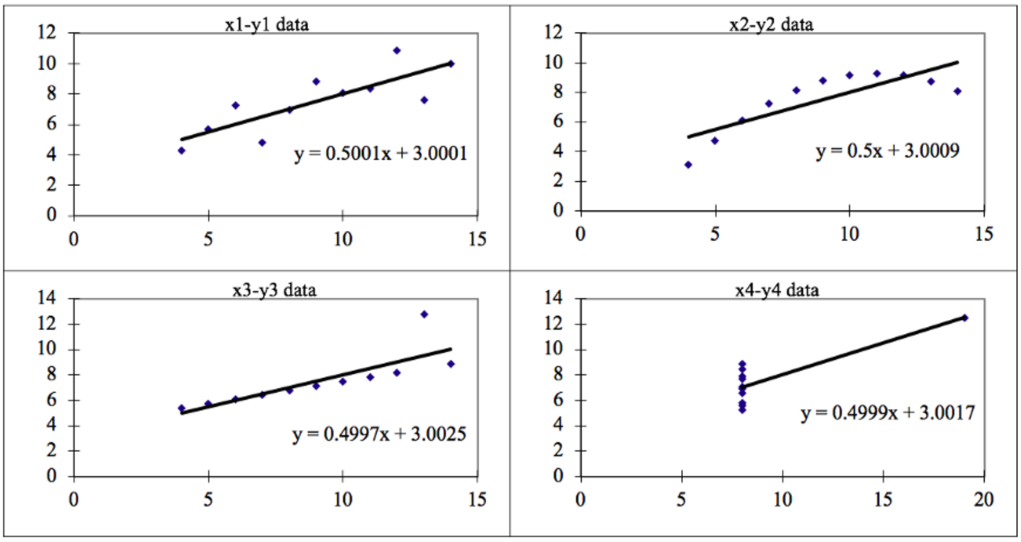


Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Anscombe's quartet highlights the importance of **plotting data to confirm the validity of the model fit**. In each panel, the Pearson correlation between the x and y values is the same, r = 0.82 (approx.). In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values.

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

Dataset 1: this **fits** the linear regression model pretty well.

Dataset 2: this **could not fit** linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

Conclusion:

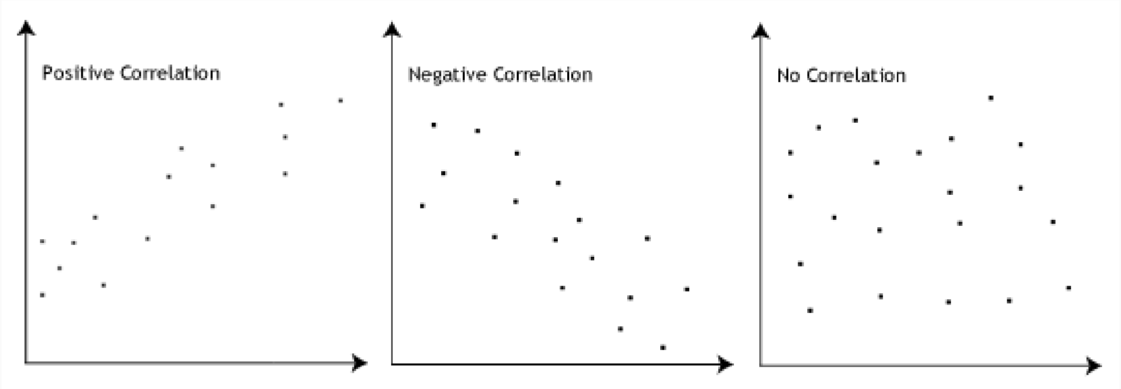
We described 4 datasets that were intentionally created in order to describe the importance of data visualisation and how and linear regression model can be fooled by the same. Therefore, it's a recommended practice to visualize all the important features in the dataset before implementing any machine learning algorithm on them.

Q. What is Pearson's R?

Answer: Pearson correlation coefficient (PCC), also referred to as **Pearson's R** 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 a normalised 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)
- 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



Q. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: **Scaling** 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.

Mostly, the collected dataset contains features that are highly varying in magnitudes, units and ranges. In case the scaling step is skipped, the algorithm only takes magnitude in account and not units which therefore leads to incorrect modelling. To solve this issue, we need to do scaling in order to bring all the variables to the same level of magnitude.

An important point to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between normalized scaling and standardized scaling:

- Normalized scaling (also known as MinMax scaling) rescales the values into a range of [0, 1] whereas, Standardized scaling rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
- Standardised scaling will affect the values of dummy variables but MinMax scaling will not.
- The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there is are extreme data point (outlier).

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

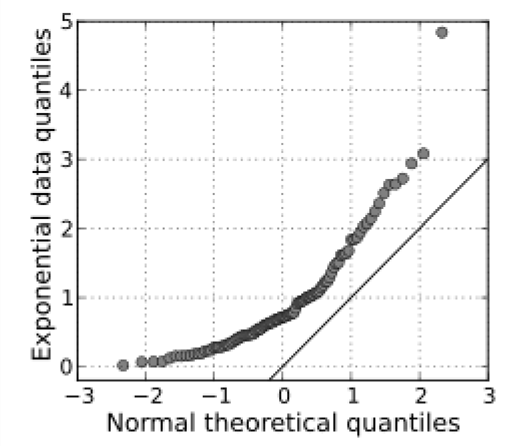
Answer: If there is perfect correlation, then VIF value turns out to be “infinity”. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which leads to 1/(1-R2) which gives the value “infinity”.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which shows an infinite VIF as well).

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

Answer: Q-Q Plots refers to **Quantile-Quantile** plots which are the plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. A quantile ranges from 0 to 100 percentile. The purpose of Q-Q plot is to figure out if the two datasets 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.

Here is a Q-Q plot showing the 45 degree 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.

- A Q–Q plot can be used as a graphical means of estimating parameters in a location-scale family of distributions.
- Also, it 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.