upGrad

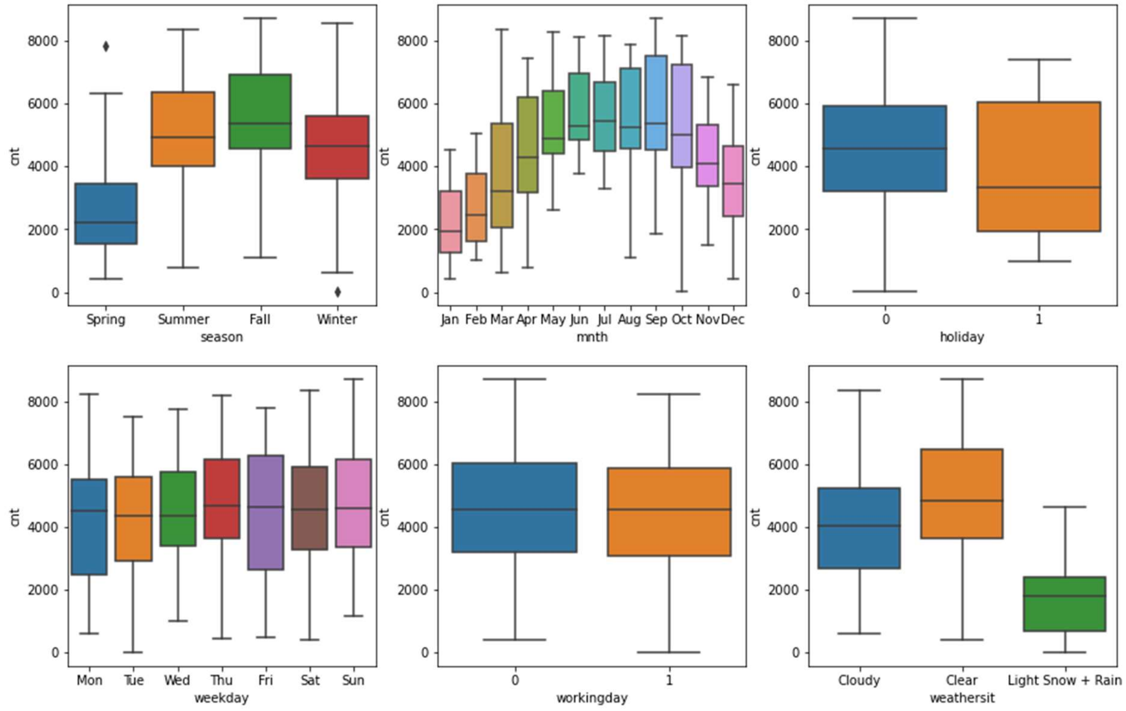
madanhm9945@gmail.com

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?**

   Here is a sub plot with 6 categorical variables being plotted and mentioned below are inferences made with this.

- **Season**: Bookings attracted more during Fall and less during spring, winter and summer have similar figure.
- **Month**: July has more business results whereas Jan has less count.
- **Holiday**: Count is reduced during holidays.
- **Weekday**: It actually looks average when median is observed, but count wise, Friday holds higher rank.
- **Working day**: No much impact whether it's a working day or non-working day, bookings are attracted.
- **Weathersit**: Clear day attracted more whereas Light snow + rain attracted less business, quite expected.

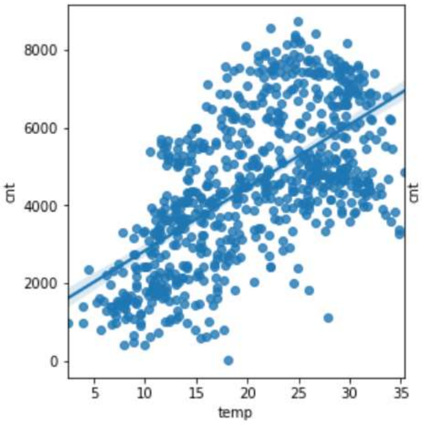


2. **Why is it important to use drop_first=True during dummy variable creation?**

So to answer this, it's a bit of logic as well as efficiency of data which is important to manage. For example, we have a variable with three components, a dummy can be created for this with 2 components dropping anyone of those, which doesn't impact on identification of any of the component.

For example: **Flats available in an apartment** – 1BHK can be represented as 00, 2BHK as 01 and 3BHK as 11. So here 1 and 0 are two values with their position in a 2 digit, identifying the type of variable which doesn't impact in any of the components.

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

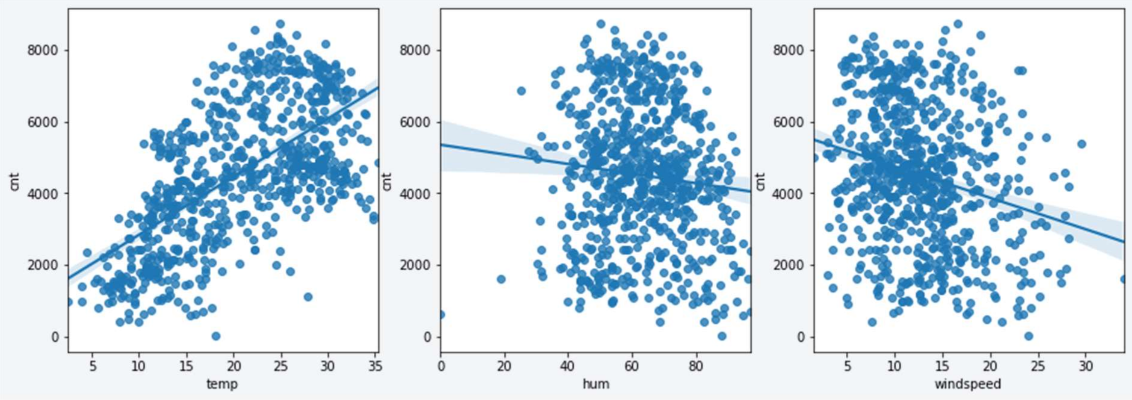


Temperature variable has highest positive correlation with the target variable count, which is fairly true, as the temperature increases, bookings increased as per the regression plot drawn here.

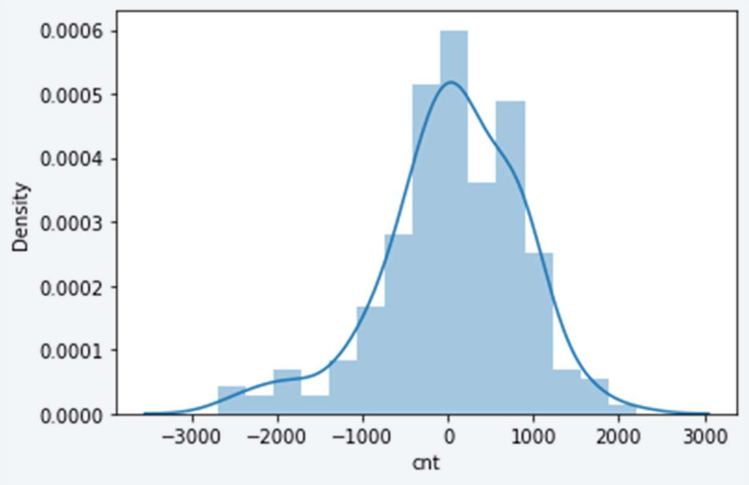
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Assumptions are validated and results are mentioned below with appropriate plots.

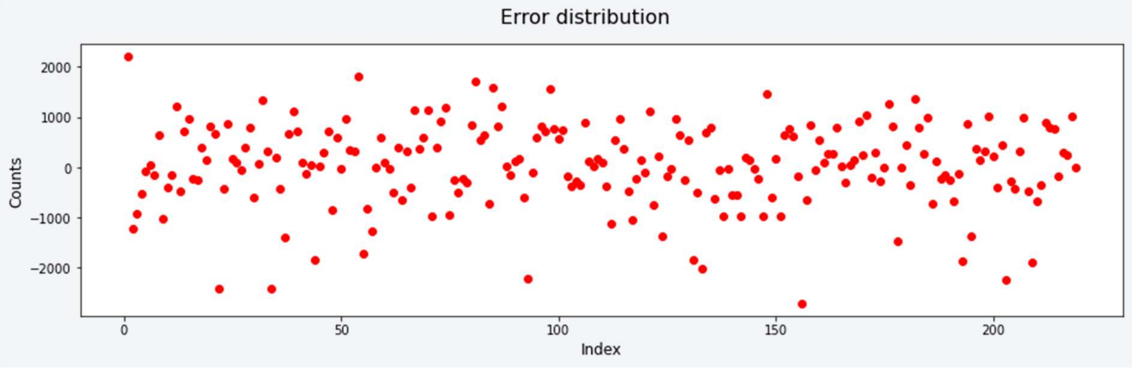
There is a *linear relationship* between X and Y:



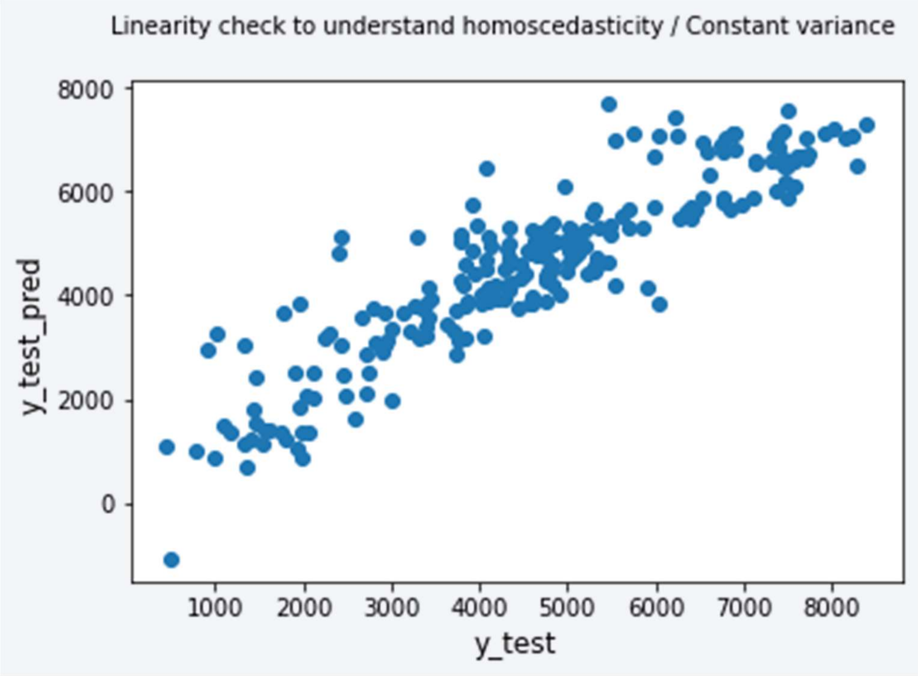
Error terms are *normally distributed* with mean zero (not X, Y):



Error terms are *independent* of each other:



Error terms have *constant variance* (homoscedasticity):



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

The top three influencing variables are as mentioned below:

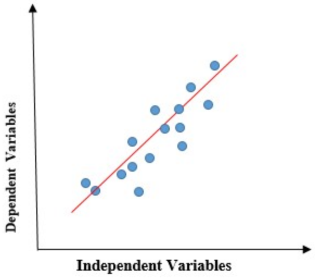
1. **Temp – 0.4094** - It is obvious as temperature increases, bike bookings increase, since people doesn't prefer to walk on sunny day.
2. **Weathersit_Light Snow + Rain: - 0.2571** – Inversely affecting the bookings, running offers preferred during this time
3. **Year: 0.2034** – Bookings increased with the progression in year

madanhm9945@gmail.com

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression 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. 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. A technical definition would be **"Linear Regression Algorithm is a machine learning algorithm based on supervised learning.".** Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.



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.

Simple linear regression uses traditional slope-intercept form, where mm and bb are the variables our algorithm will try to "learn" to produce the most accurate predictions. xx represents our input data and yy represents our prediction.

y=mx+b

x = Independent variable

y = Dependent variable

m = Slope of the line (d2-d1 / c2-c1) where d and care points on a straight line

b = intercept of the line

Multi Linear regression:

A more complex, multi-variable linear equation might look like this, where ww represents the coefficients, or weights, our model will try to learn.

$$y_i = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_p * x_n + \epsilon$$

y=dependent variable

xi=explanatory variables

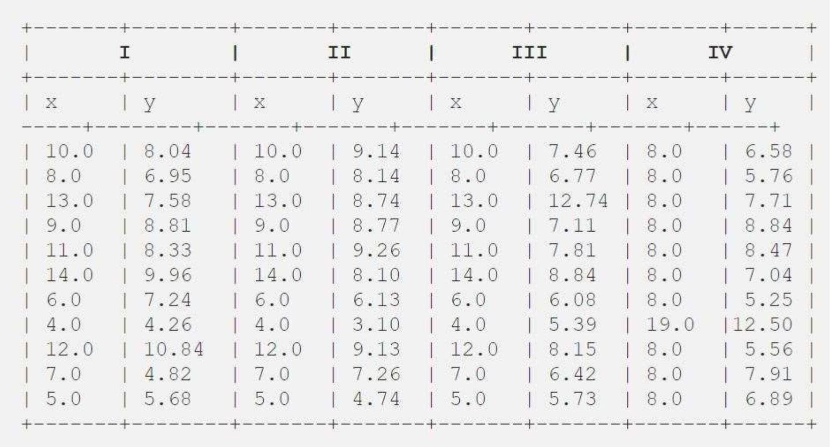
β0=y-intercept (constant term)

βp=slope coefficients for each explanatory variable

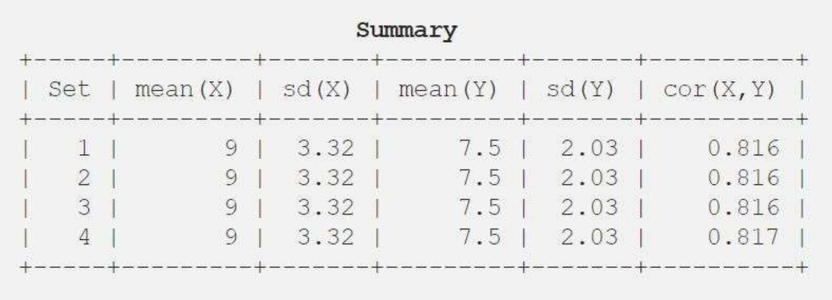
ϵ=the model's error term (also known as the residuals)

2. Explain the Anscombe’s quartet in detail.

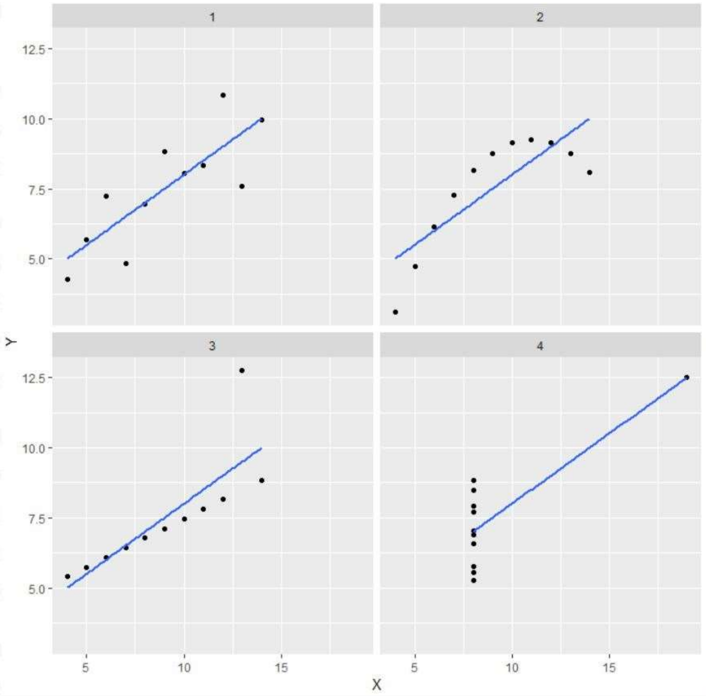
Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```



```
                         Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```



It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where,
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

A data Pre-Processing method which is applied to independent variables to normalize the data within a particular range which inturn helps in speeding up the calculations in an algorithm.

Eventually, collected data set contains features highly varying in magnitudes, units and range. Scaling is performed to consider the units otherwise magnitude will be taken into account which is an error in interpretation of model. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. **Scaling should only affect the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.
- **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. Hence, If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. We need to drop one of the variables from the dataset which is causing this perfect multicollinearity in order to avoid perfect correlation and infinite VIF.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

In python, statsmodels.api provides **qqplot** and **qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.

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. 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.